An Information-Theoretic Approach to Methylation Data Analysis

John Goutsias

Whitaker Biomedical Engineering Institute
The Johns Hopkins University
Baltimore, MD 21218

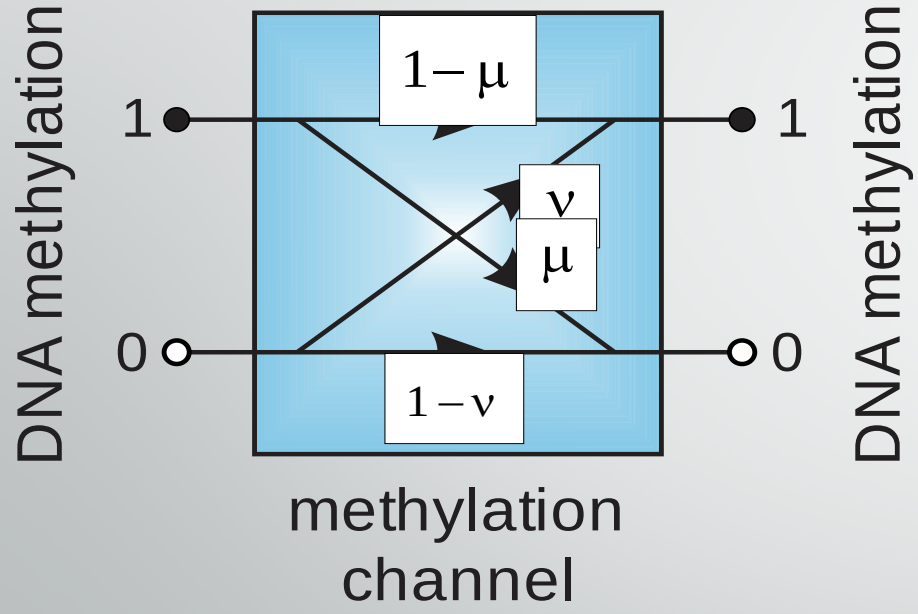
Objective

- Present an information-theoretic approach to methylation data analysis.
- This leads to:
  - Employing the Shannon entropy to quantify epigenetic stochasticity.
  - Predicting large-scale chromatin organization (compartments A/B).
  - Predicting smaller-scale chromatin organization (TADs).
  - Using an information-theoretic distance to quantify epigenetic discordance.
  - Developing a sensitivity analysis approach to quantify environmental influences on epigenetic stochasticity.

Methylation Maintenance

μ : demethylation probability

ν : methylation probability

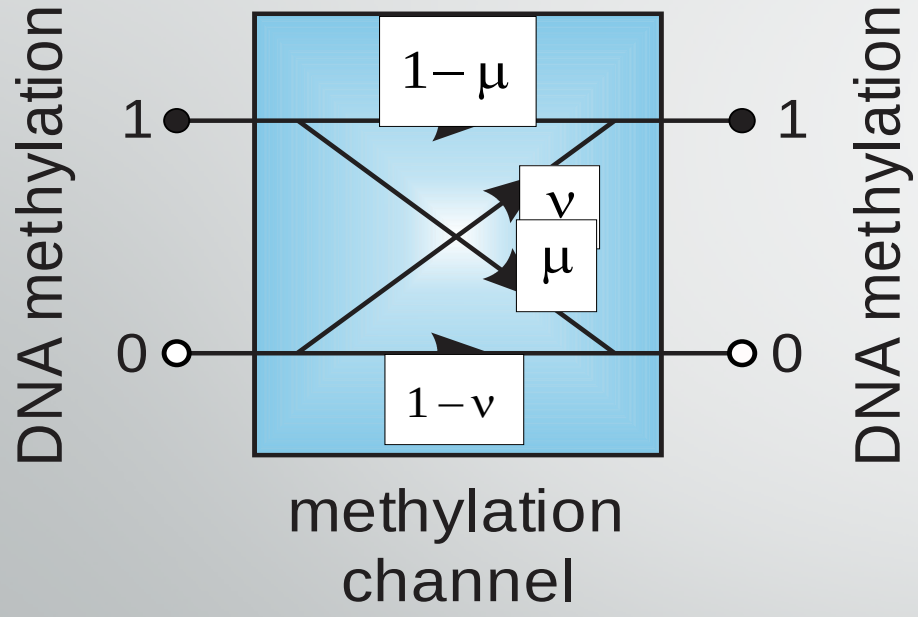


$$P_{out}(0) = (1-\nu)P_{in}(0) + \mu P_{in}(1)$$
$$P_{out}(1) = \nu P_{in}(0) + (1-\mu)P_{in}(1)$$

Methylation Maintenance

μ : demethylation probability

ν : methylation probability



$$P_{out}(0) = (1-\nu)P_{in}(0) + \mu P_{in}(1)$$
$$P_{out}(1) = \nu P_{in}(0) + (1-\mu)P_{in}(1)$$

- Use **methylation channels** at each CpG site (noisy binary communication channels).
- Must estimate probabilities μ and ν from data
- Not possible !
- At steady-state, $P_{out} = P_{in}$ and

$$\lambda = \frac{\nu}{\mu} = \frac{P_{in}(1)}{1-P_{in}(1)}$$

Information Capacity of a Methylation Channel

- Maximum average information that can be conveyed during a maintenance step:

$$C = \max_{P(1)} \{ I(X_{out}, X_{in}) \}$$

mutual information

- **Mutual information:** a measure of mutual dependence between input and output methylation.
- Available formula for capacity, but depends on probabilities μ and ν.
- Approximate formula can be derived that depends **only** on $\lambda = \nu / \mu$.

Probability of Transmission Error

$$\pi = \Pr[\text{output methylation} \neq \text{input methylation}]$$

- Quantifies fidelity of methylation transmission through a methylation channel.
- Depends on probabilities μ and ν.
- An approximate formula can be derived that depend **only** on $\lambda = \nu / \mu$.

Energy Dissipation

- Correct transmission of the methylation state requires work.
- This consumes free energy that is dissipated to the surroundings in the form of heat.
- The minimum dissipated energy is related (for some systems) to the probability of error

$$E \sim k_B T \ln \pi$$

- **Relative dissipated energy:**

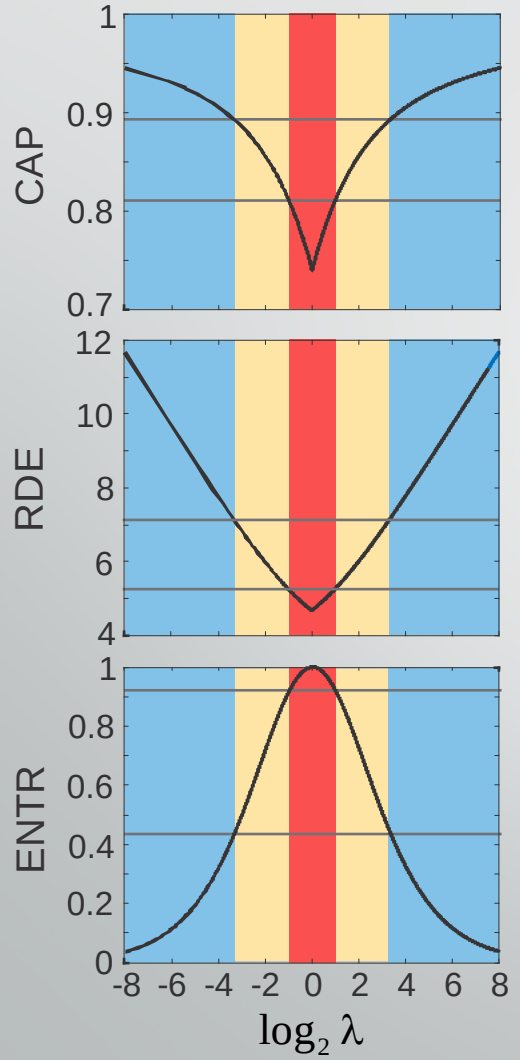
$$\varepsilon - \frac{E}{E_{max}} = -\log_2 \pi$$

Methylation Stochasticity

- We can characterize methylation stochasticity at a CpG site using the **Shannon entropy:**

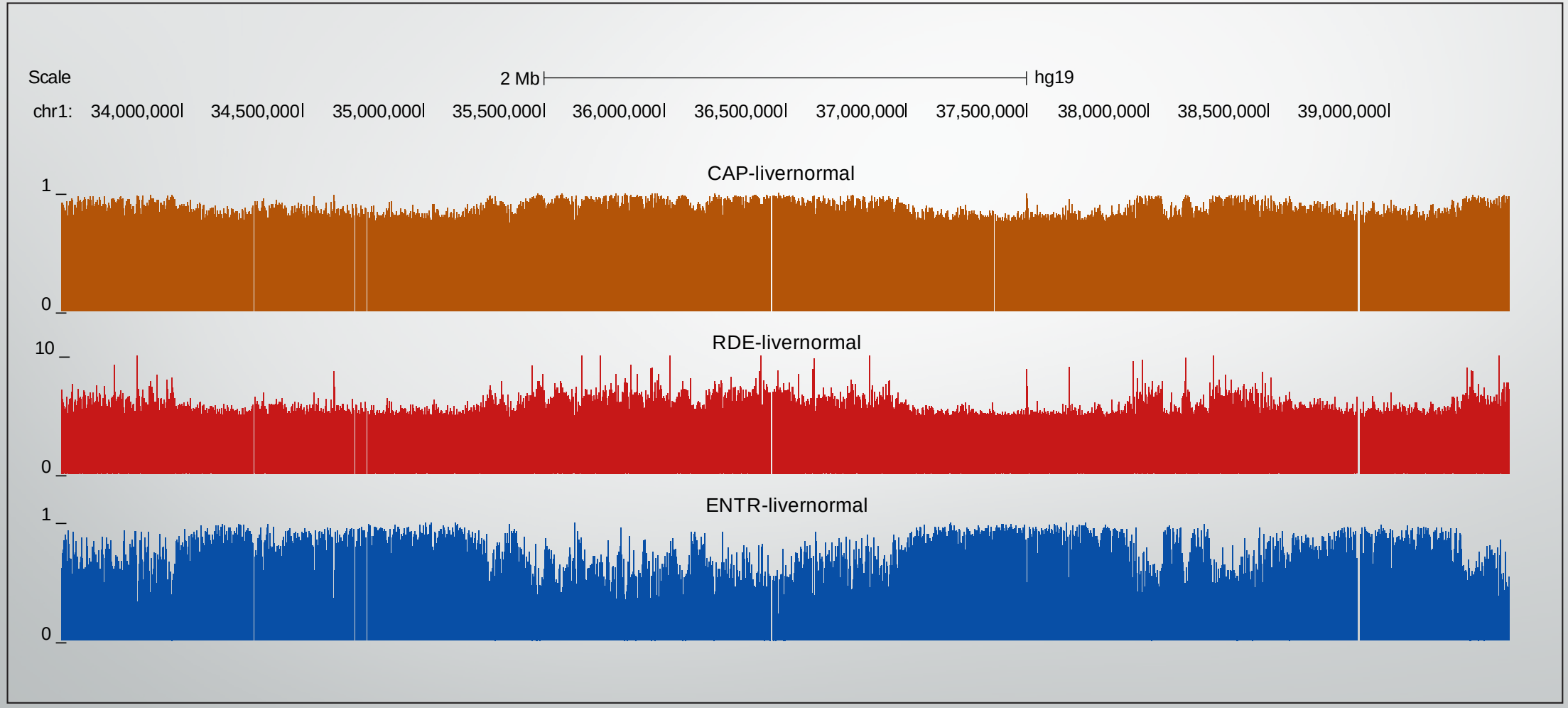
$$S = -P(0)\log_2 P(0) - P(1)\log_2 P(1)$$

Capacity, Relative Dissipated Energy, Entropy

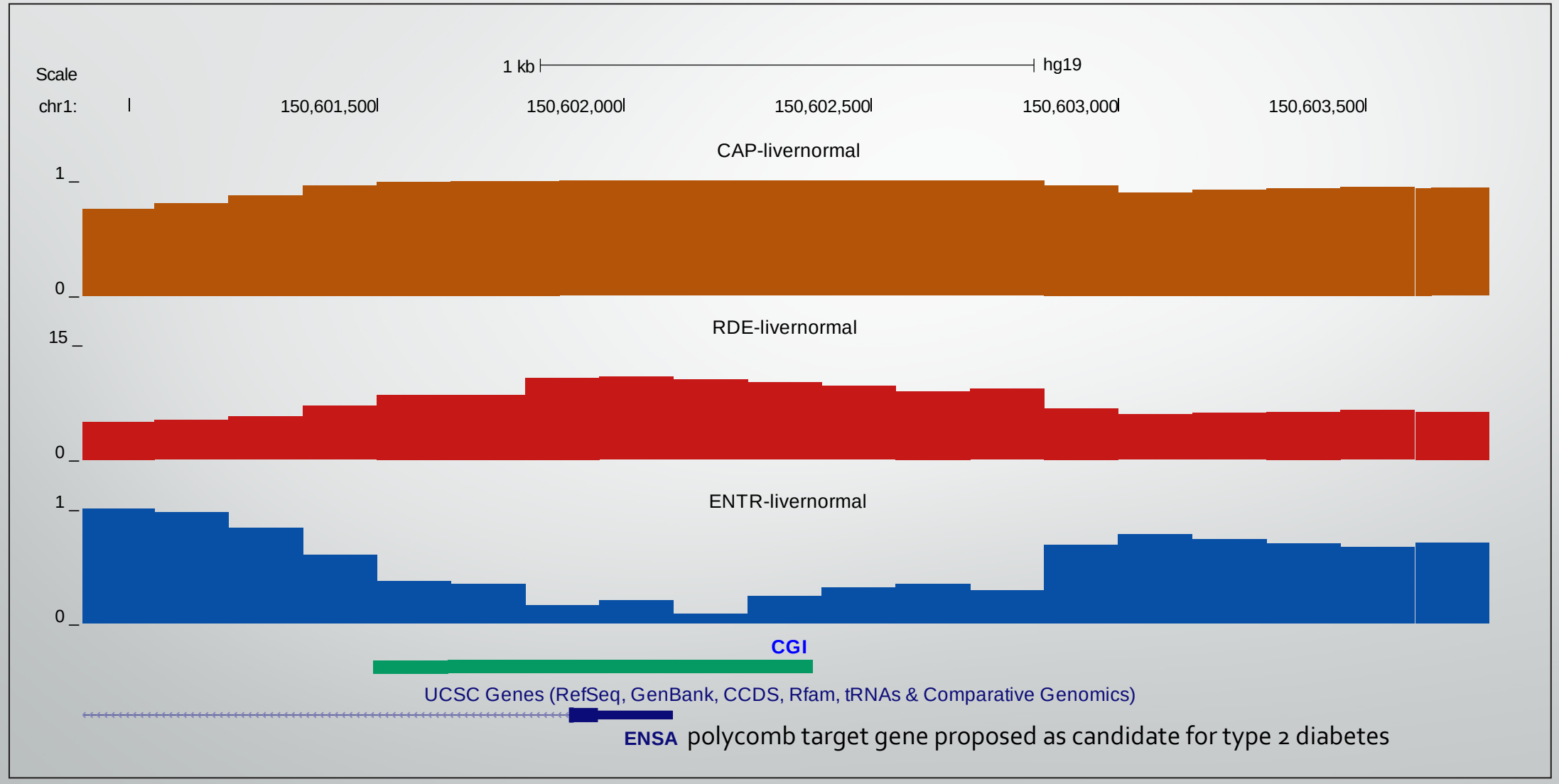


- **Reliable transmission** of methylation information within critical regions of the genome is facilitated by **high capacity** methylation channels that result in **low methylation stochasticity** at the cost of **high energy consumption**.
- **Unreliable methylation transmission** within other regions of the genome due to **low capacity** methylation channels that **consume less energy** but produce **higher levels of methylation stochasticity**.

Capacity, Relative Dissipated Energy, Entropy

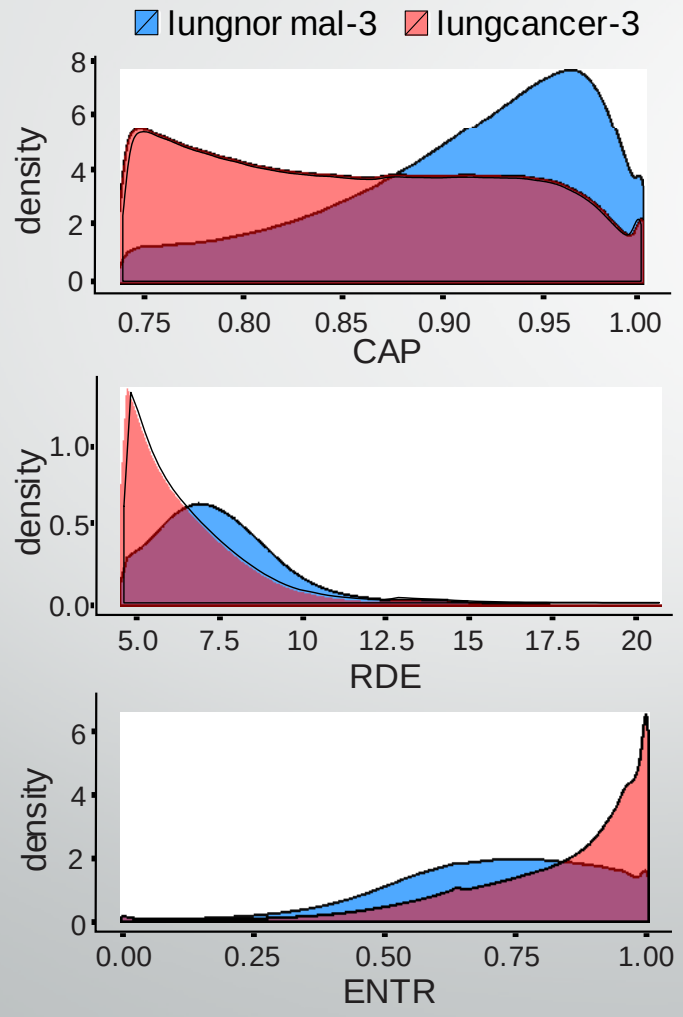


Capacity, Relative Dissipated Energy, Entropy

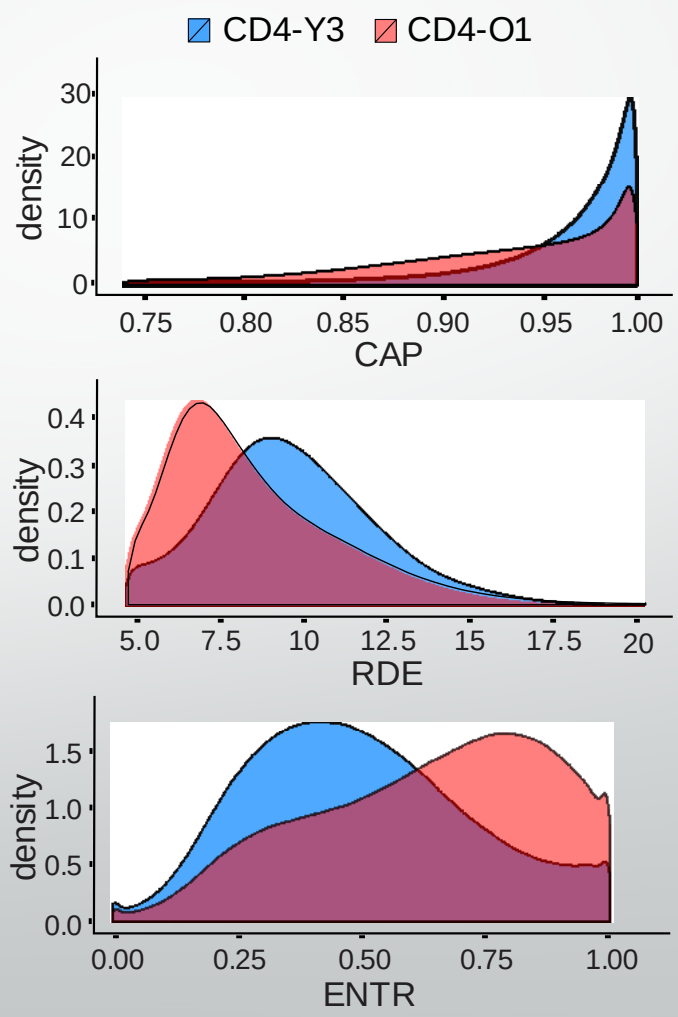


Capacity, Relative Dissipated Energy, Entropy

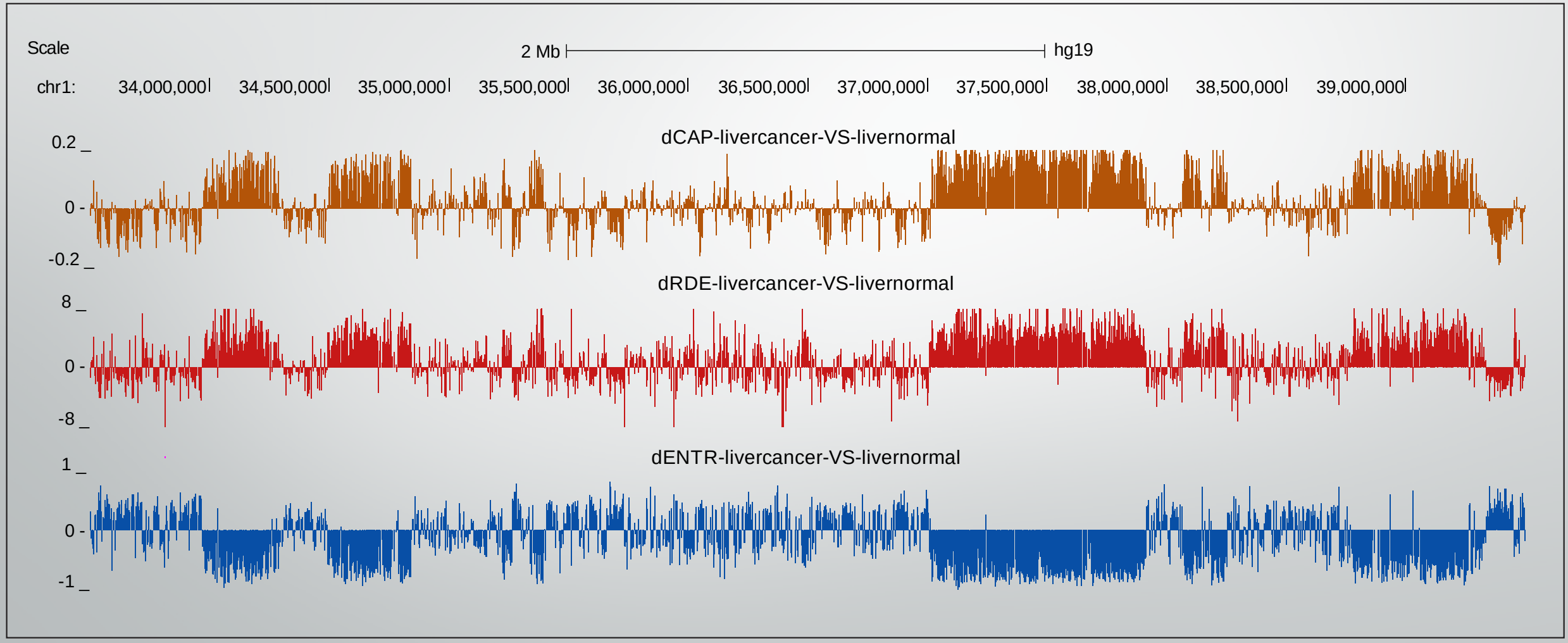
NORMAL/CANCER



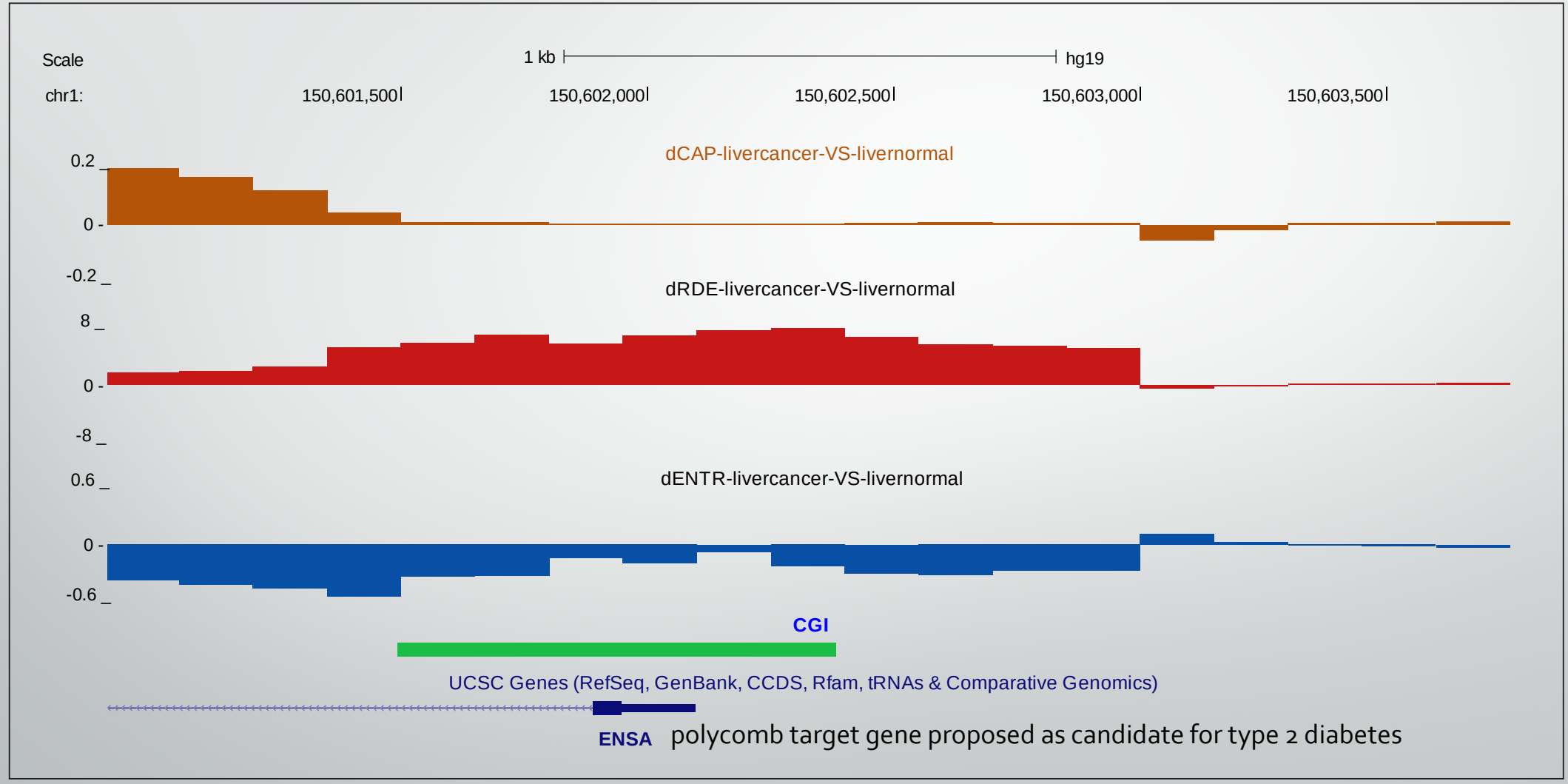
YOUNG/OLD



Capacity, Relative Dissipated Energy, Entropy



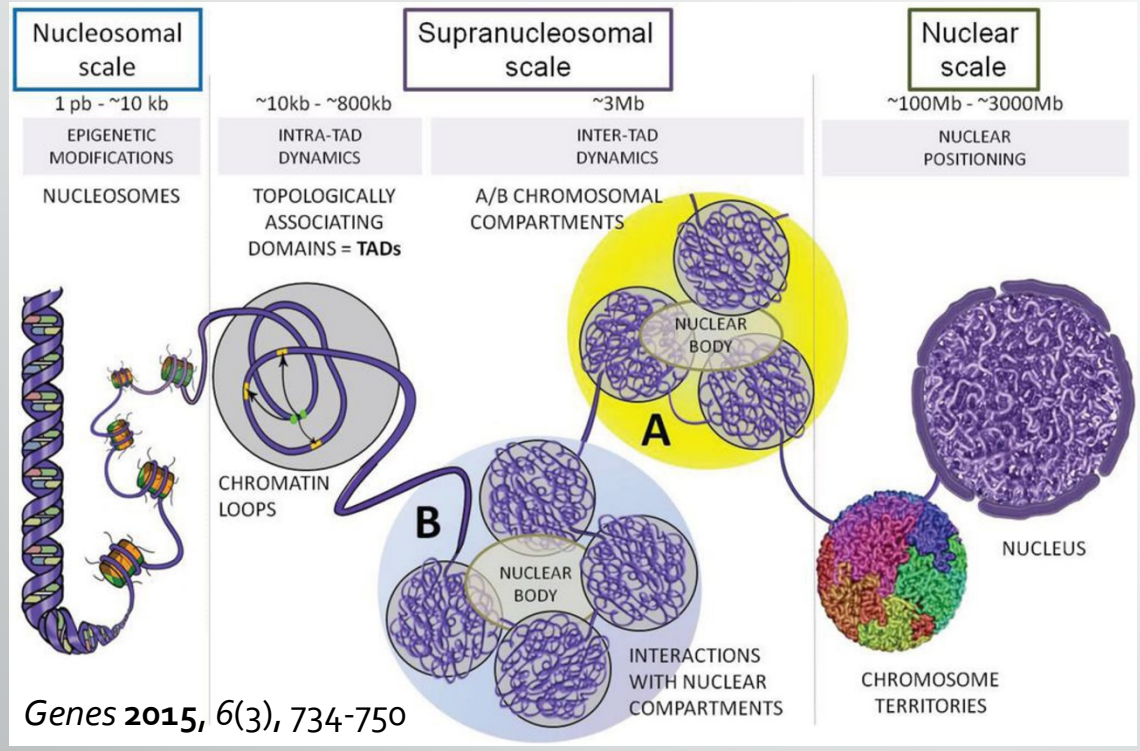
Capacity, Relative Dissipated Energy, Entropy

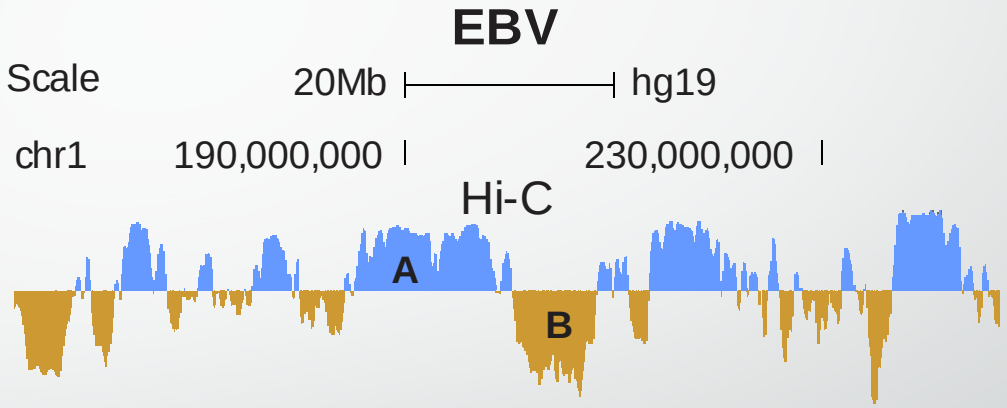


polycomb target gene proposed as candidate for type 2 diabetes

Information-Theoretic Prediction of Chromatin Organization

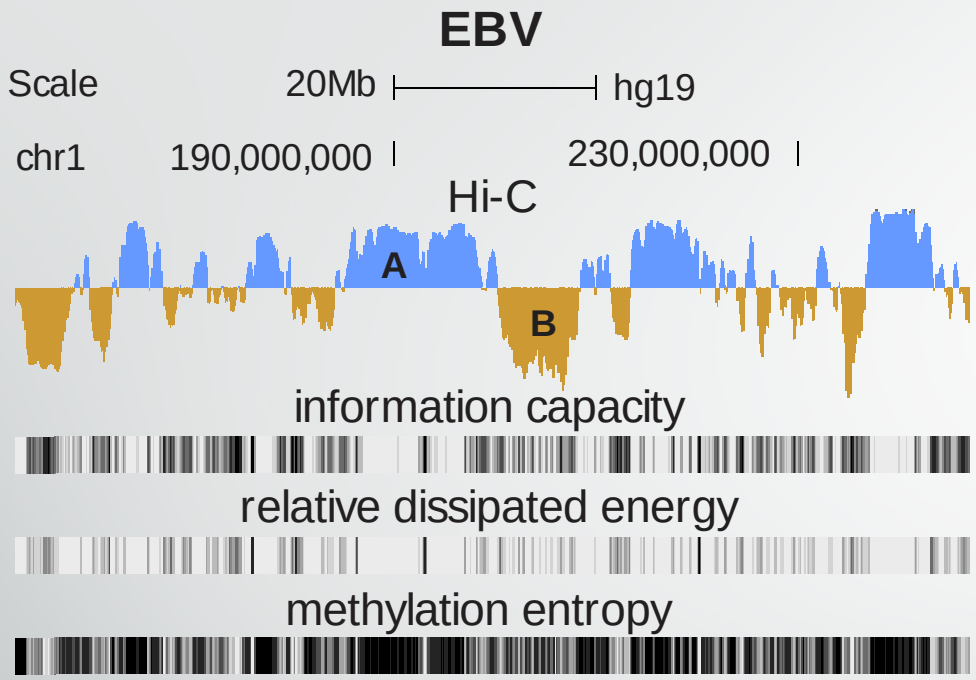
- Chromatin is organized in cell-type specific compartments A and B.

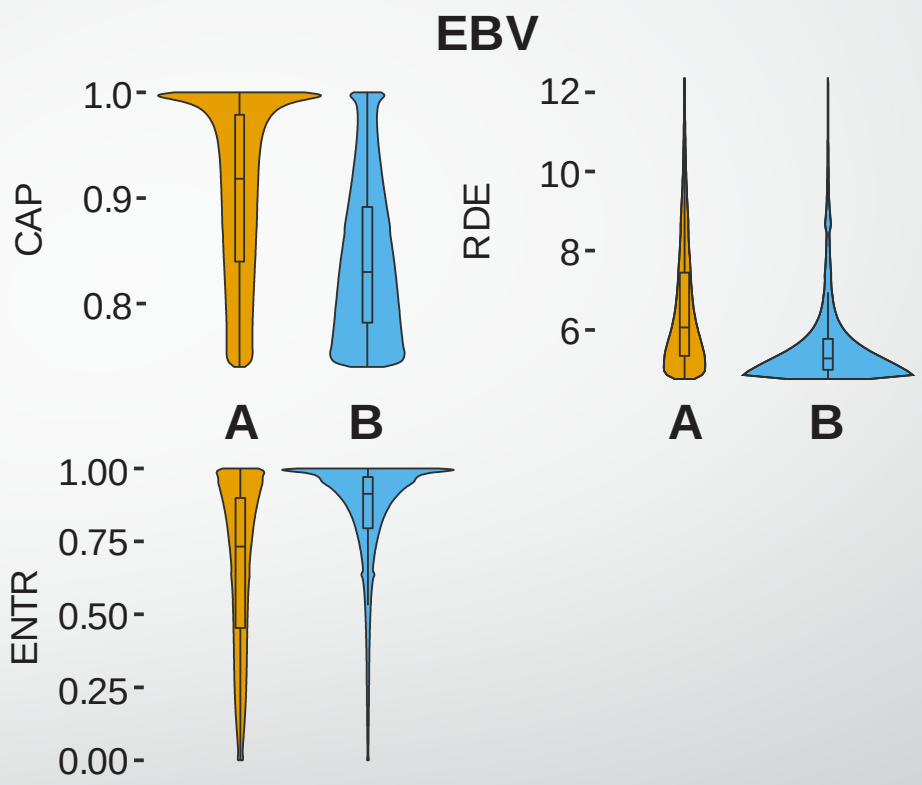




- **Compartment A**: associated with gene-rich transcriptionally active open chromatin.
- **Compartment B**: associated with gene-poor transcriptionally inactive chromatin.

Information-Theoretic Prediction of Chromatin Organization





- Enrichment of **low information capacity**, **low relative dissipated energy**, and **high entropy** within compartment B !
- Opposite true for compartment A !

Information-Theoretic Prediction of Chromatin Organization

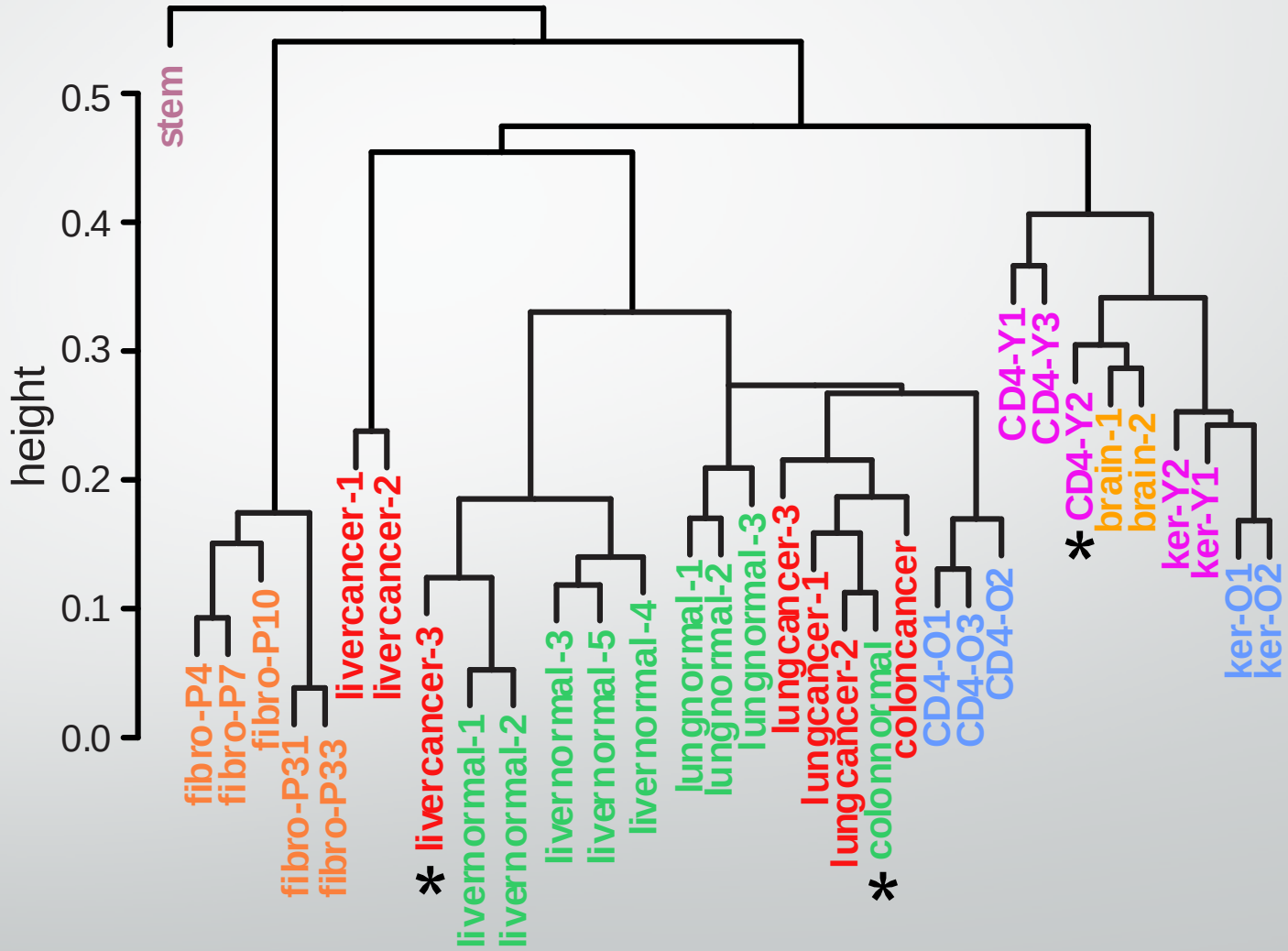
- **Random forest regression**, using information-theoretic quantities as features, learns the informational structure of compartments A/B from available "ground-truth" data.

- Achieves reliable prediction: 0.82 cross-validated average correlation and 91% average agreement between predicted and "true" A/B signals.

- **A small number of local information-theoretic properties of methylation maintenance can be highly predictive of large-scale chromatin organization !**

Information-Theoretic Prediction of Chromatin Organization

- Clustered 34 samples using the net percentage of A/B switching as a dissimilarity measure.
- 31/34 samples grouped in a biologically meaningful manner !



Quantifying Epigenetic Stochasticity

- Source of epigenetic stochasticity: Transmission of epigenetic information during maintenance through low capacity methylation channels that consume low levels of free energy.
- Quantify epigenetic stochasticity using the concept of **Shannon entropy**.
- To account for methylation dependence within a genomic region, must consider the ***joint*** Shannon entropy

$$H(X_1, X_2, \ldots, X_N) = -\sum_{x_1, x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) \log_2 p(x_1, x_2, \ldots, x_N)$$

$$X_n = \begin{cases} 1, & n\text{-th CpG site is methylated} \\ 0, & n\text{-th CpG site is unmethylated} \end{cases}$$

- Note that

$$H(X_1, X_2, \ldots, X_N) \leq H(X_1) + H(X_2) + \cdots + H(X_N)$$

Methylation Level

- Partition the genome into non-overlapping **genomic units** of 150 bp each (determines resolution of analysis).
- Instead of focusing on **methylation patterns** within a genomic unit, we quantify methylation using the **methylation level**

$$L = \frac{1}{N}\sum_{n=1}^{N} X_n$$

$$X_n = \begin{cases} 1, & n\text{-th CpG site is methylated} \\ 0, & n\text{-th CpG site is unmethylated} \end{cases}$$

- Compute the probability distribution $P(l)$ of the methylation level from the Ising model.

Mean Methylation Level

- The **mean methylation level** (MML) is the average of the methylation means at the CpG sites within a genomic unit

$$\mathrm{E}[L] = \frac{1}{N}\sum_{n-1}^{N}\mathrm{E}[X_n]$$

Normalized Methylation Entropy

- Quantify epigenetic stochasticity within each genomic unit using the **normalized methylation entropy** (NME)

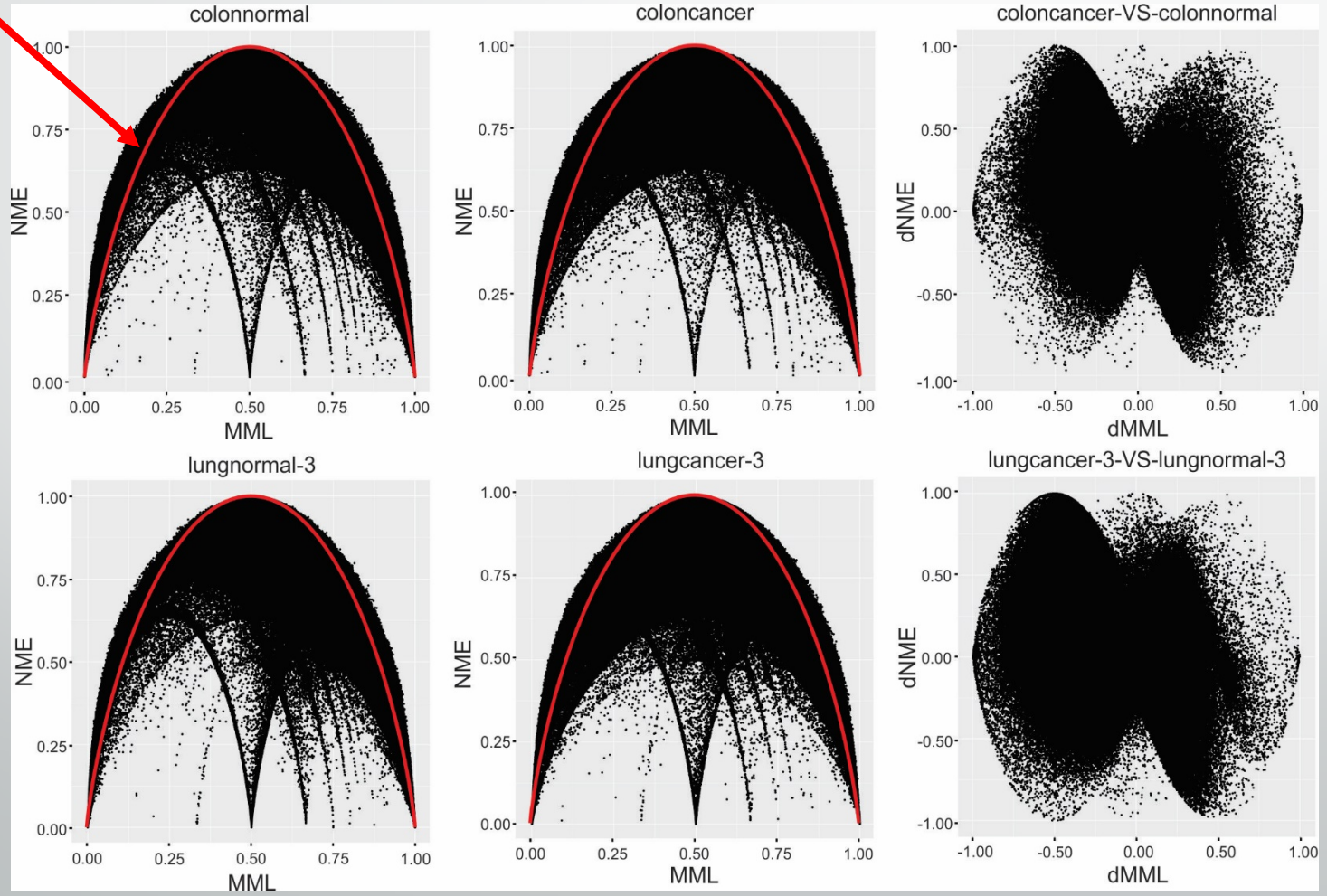
$$h = -\frac{1}{\log_2(N+1)} \sum_l P(l) \log_2 P(l)$$

- Ranges between 0 and 1
  - 0: fully ordered state
  - 1: fully disordered state

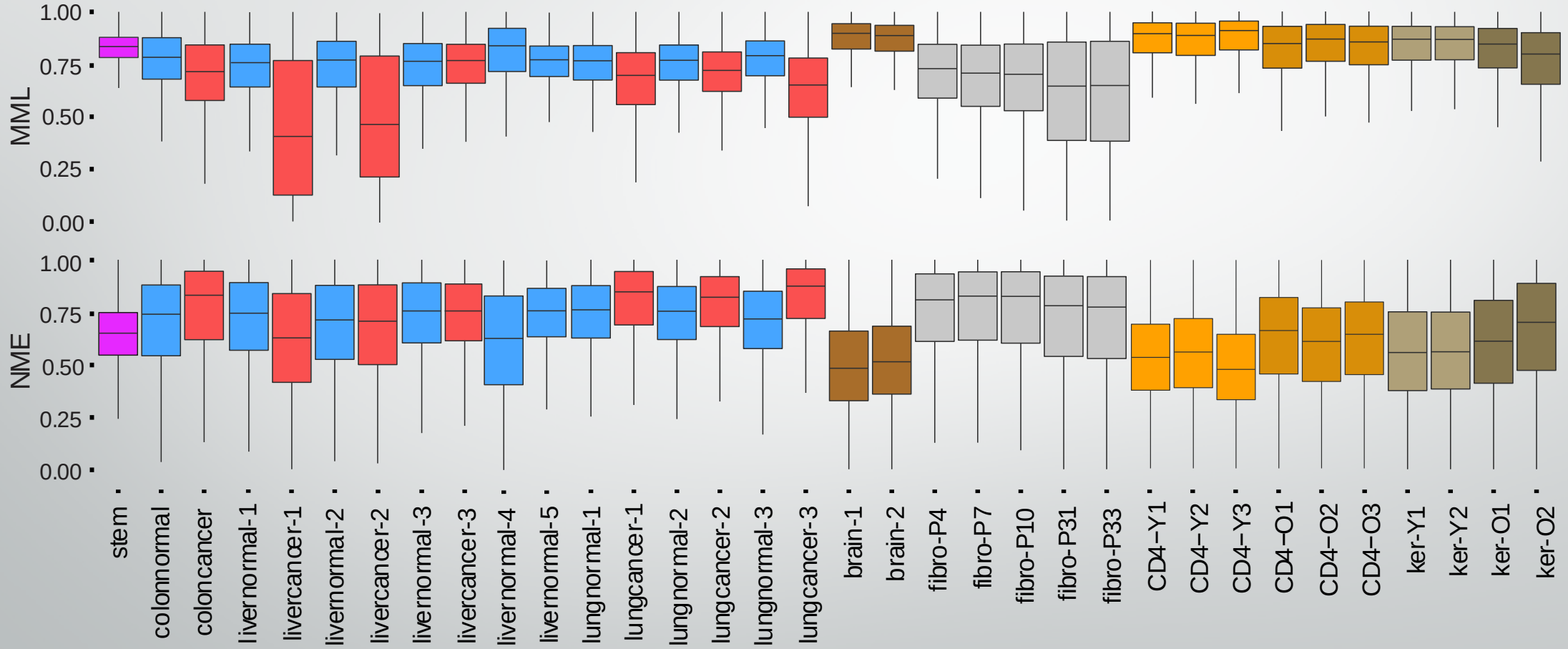
Normalized Methylation Entropy

MML not predictive of NME !

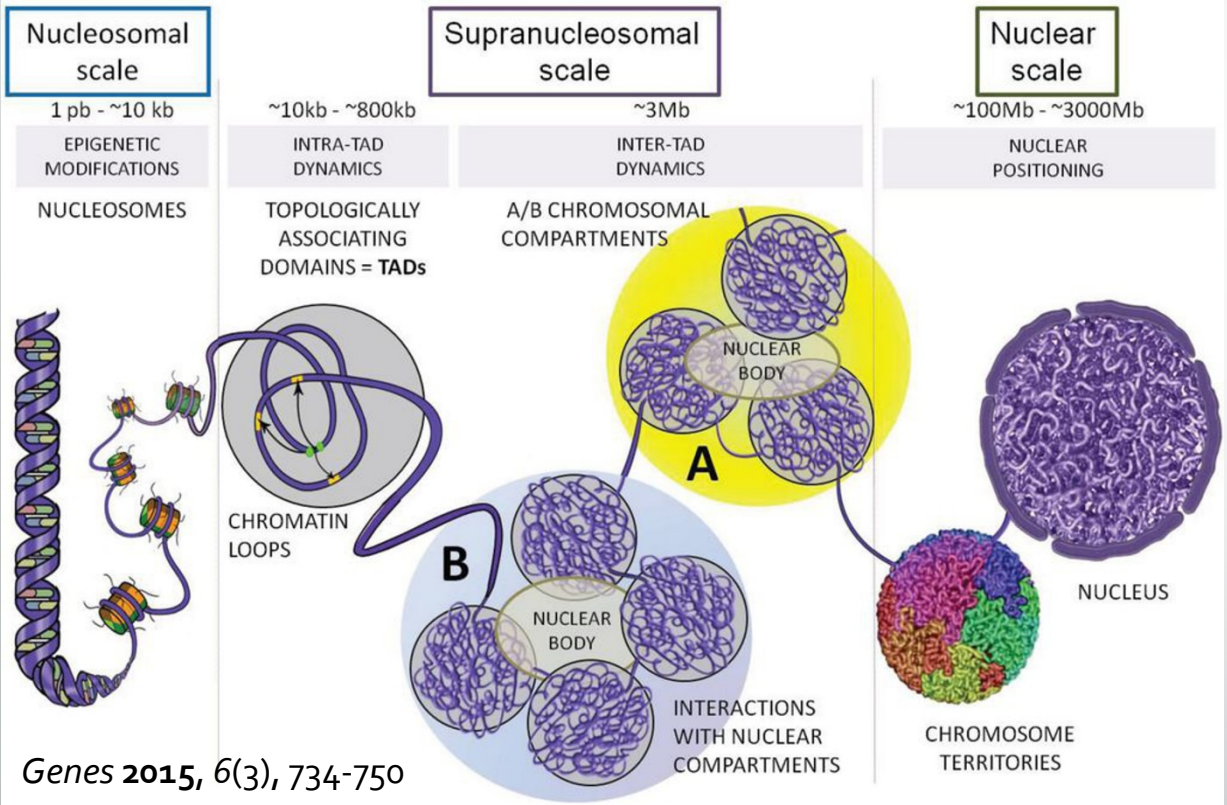
$$\text{entropy} = -\text{mean} \times \log_2(\text{mean}) - (1-\text{mean})\log_2(1-\text{mean})$$



Genome-wide MML and NME Distributions



Topologically Associating Domains

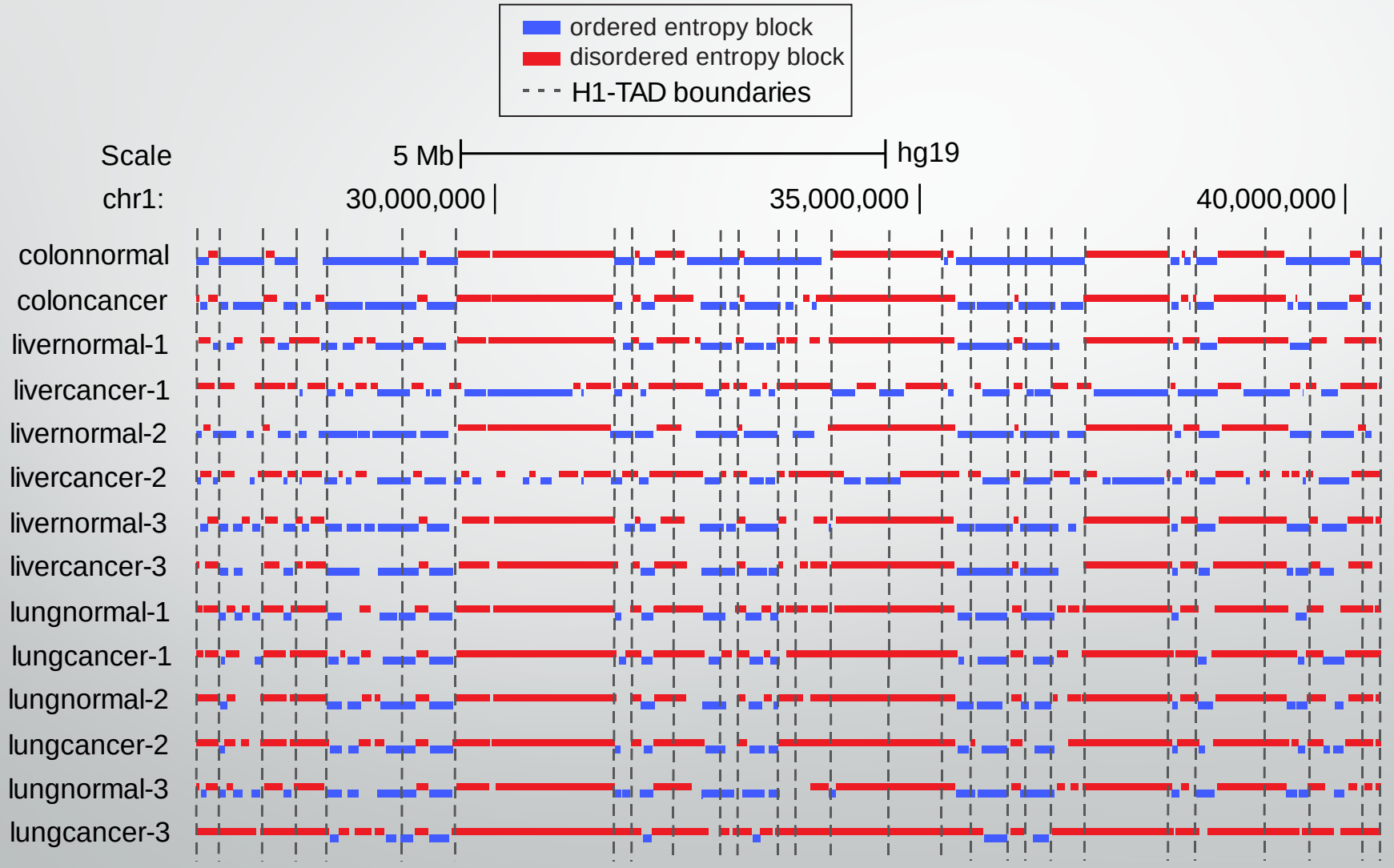


Genes **2015,** *6*(3), 734-750

- **Topologically associating domains** (TADs): highly conserved structural features of the genome across tissue types and species.
- Loci within TADs tend to interact frequently with each other.
- Less frequent interactions take place between loci within adjacent TADs.

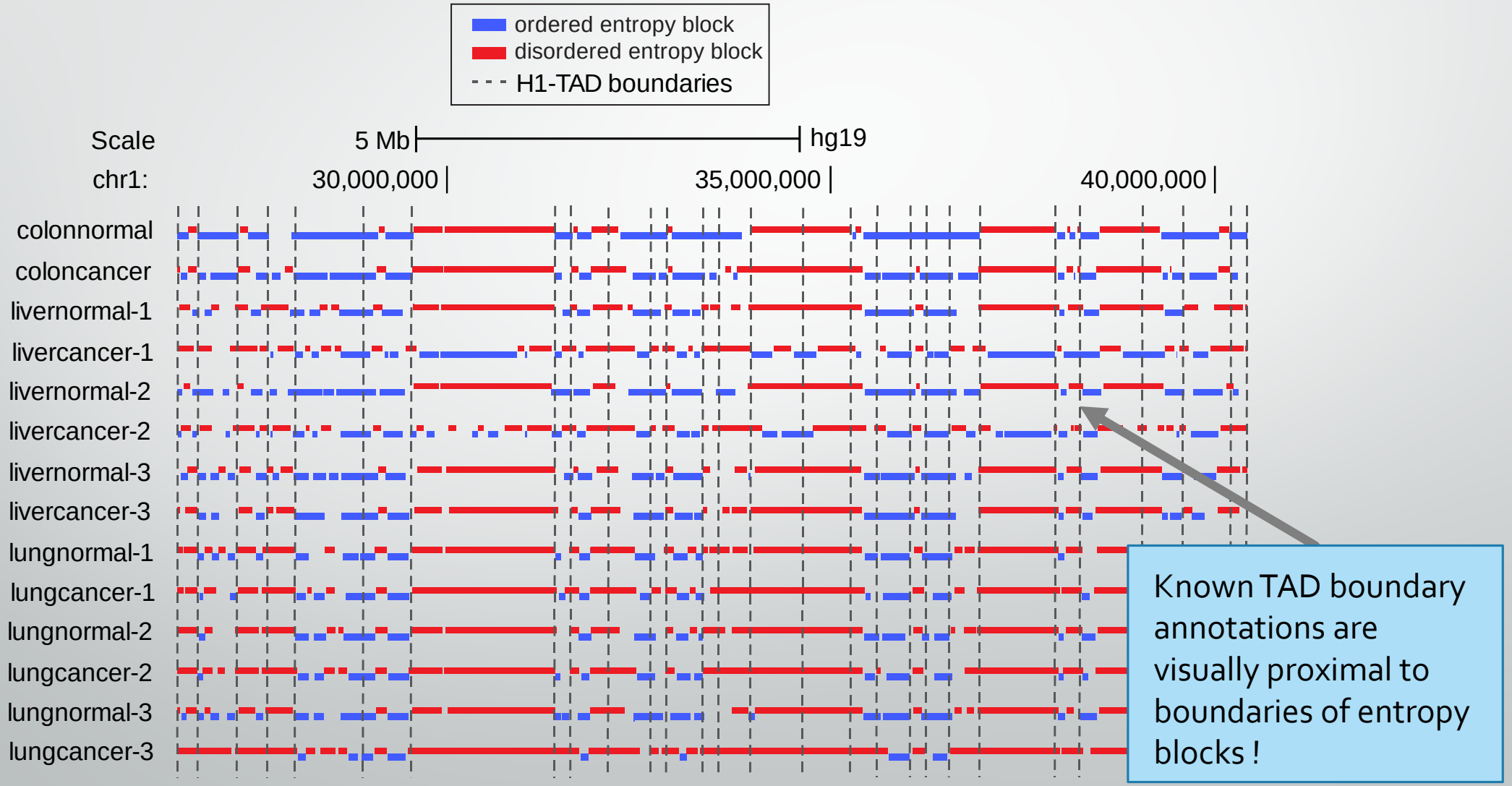
Entropy Blocks

Large genomic regions of consistently low or high NME values.

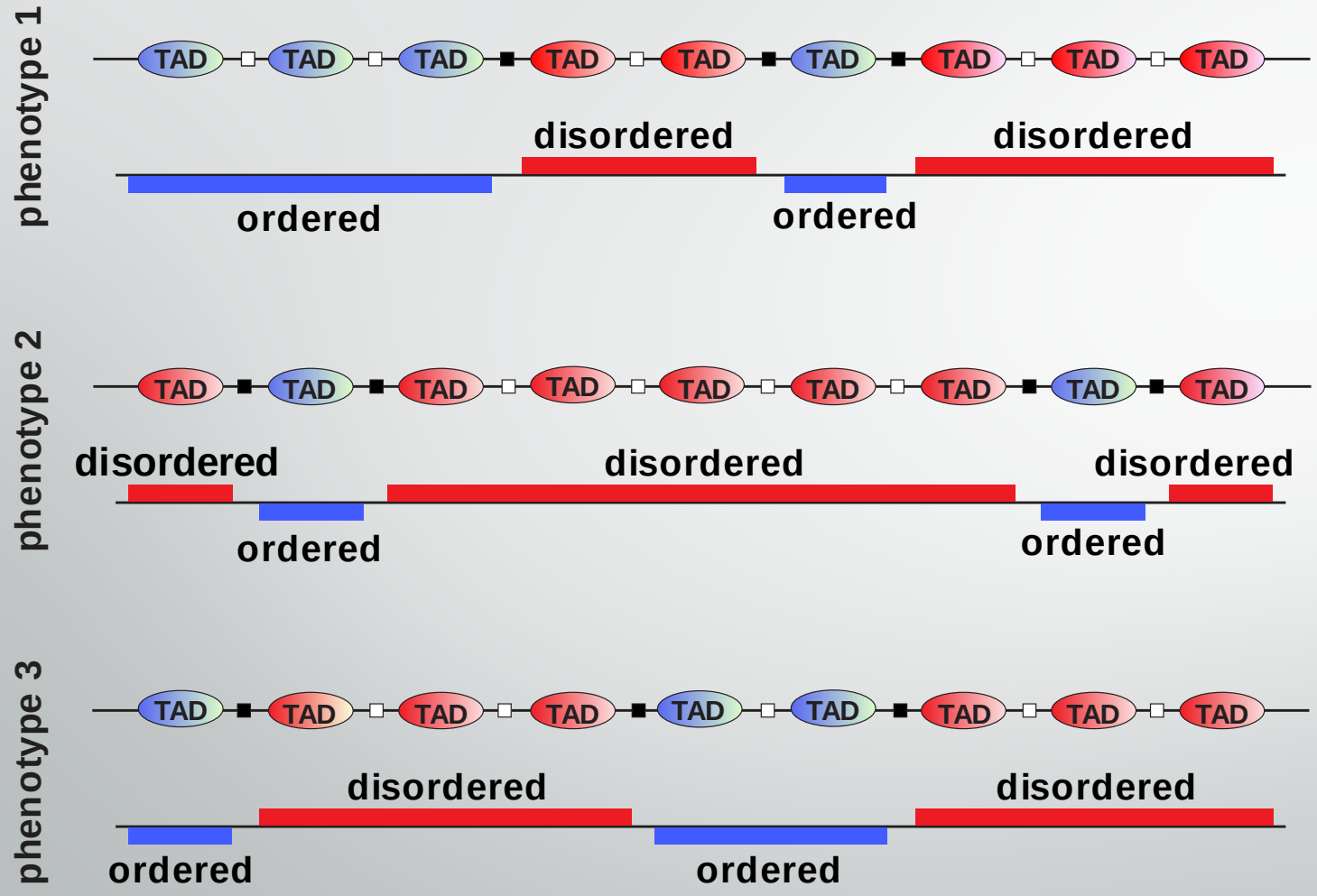


Entropy Blocks

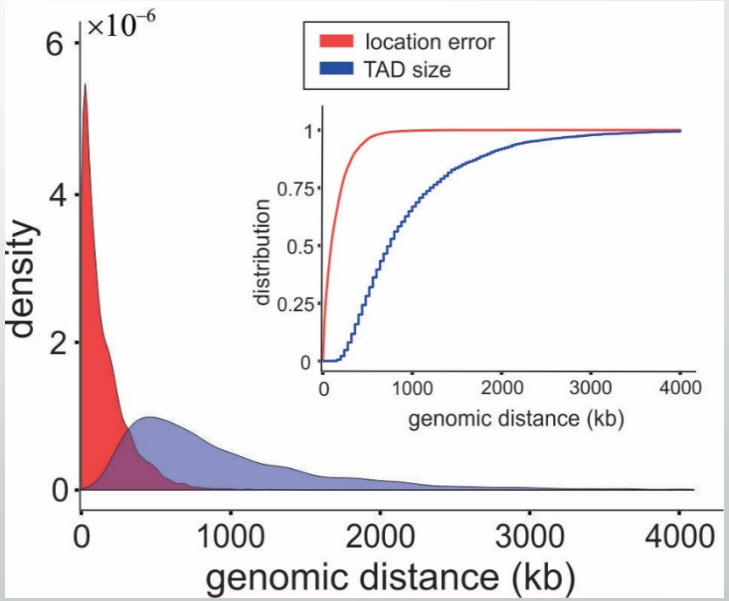
Large genomic regions of consistently low or high NME values.



Identification of TAD boundaries

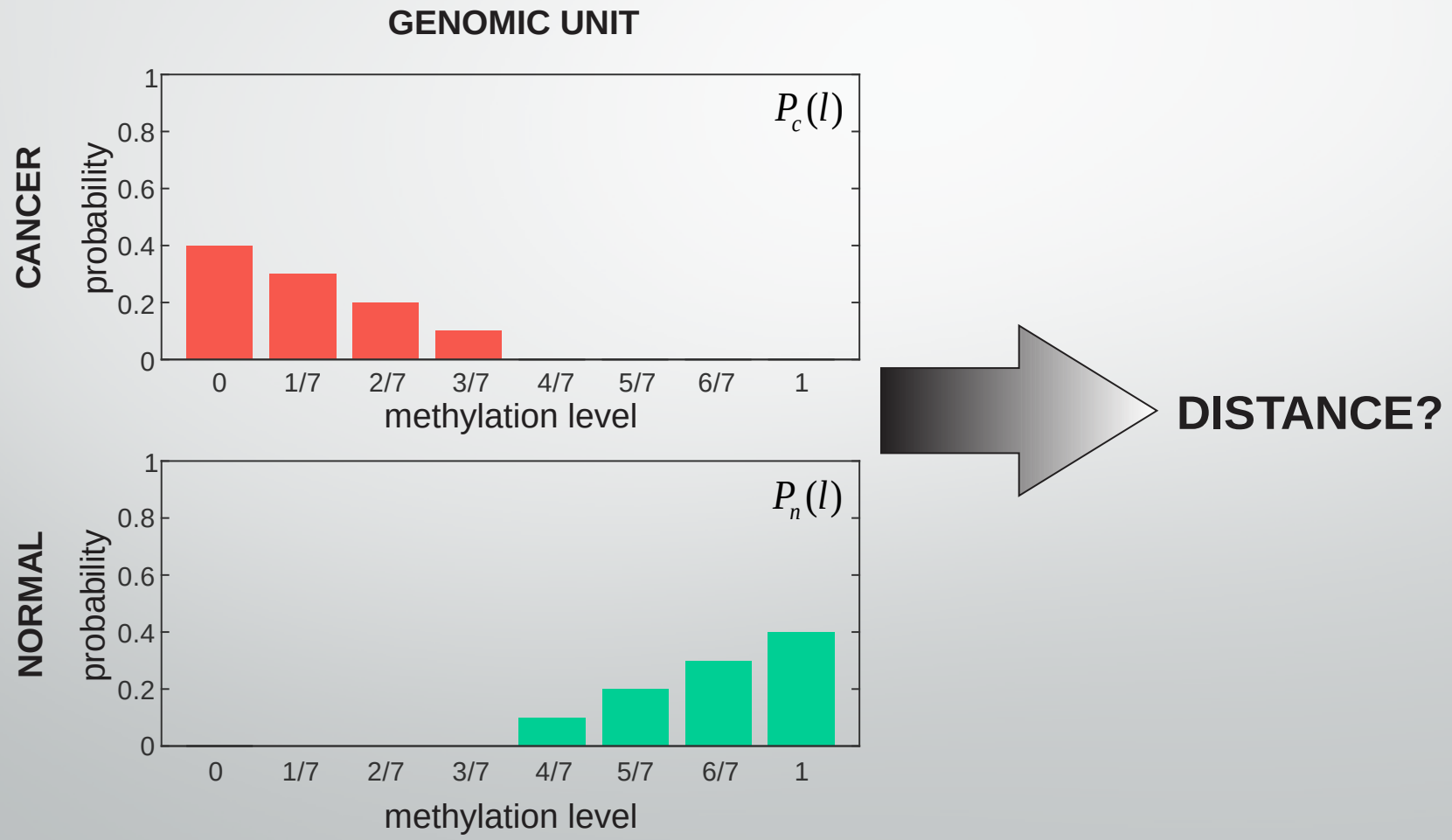


TADs are associated with consistently low or high NME values, and can be accurately identified from entropy blocks.

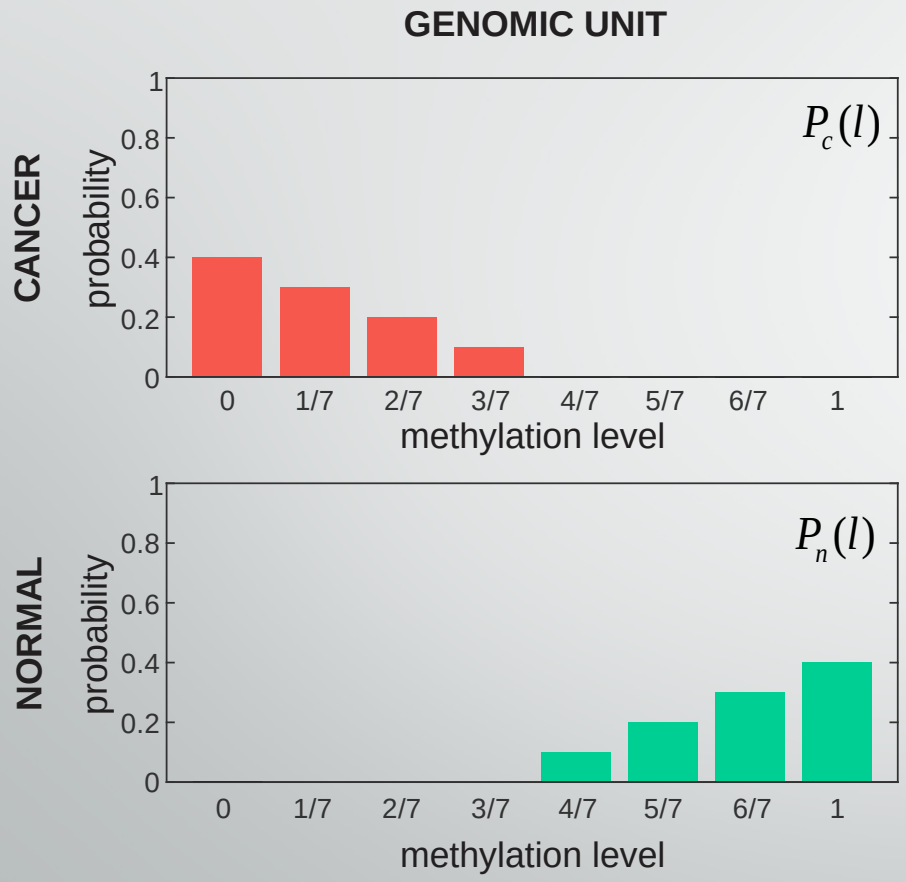


Quantifying Epigenetic Discordance

- Evaluate differences in probability distributions of methylation levels within genomic units using a distance metric.



Jensen-Shannon Distance

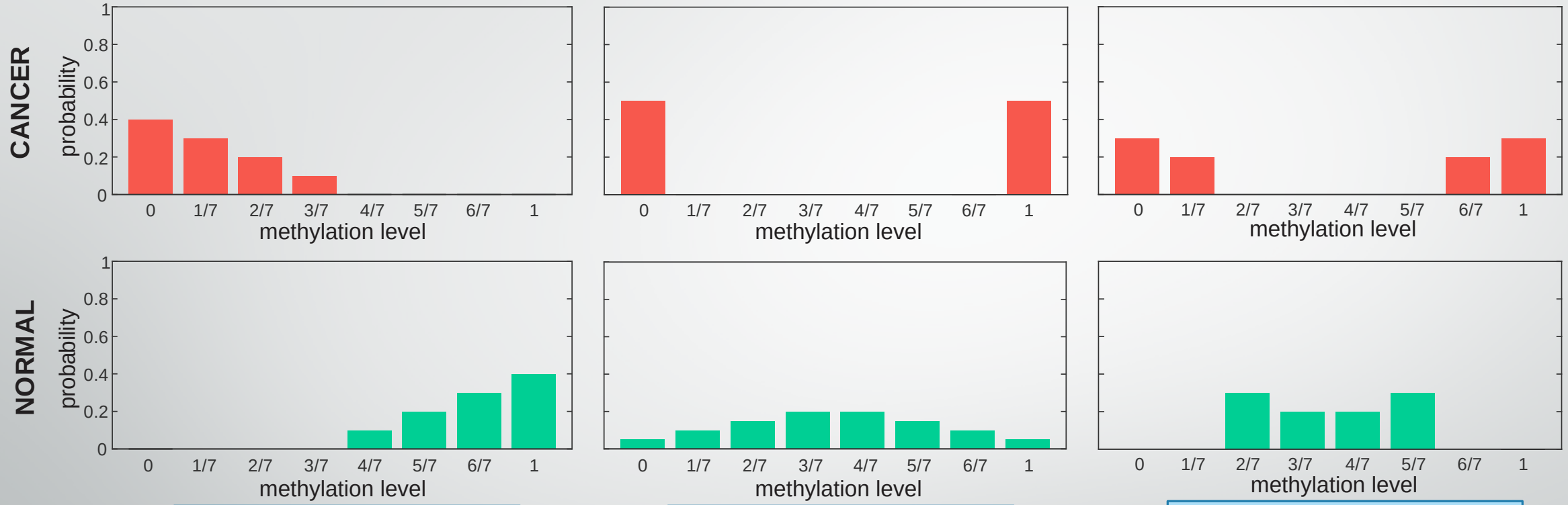


$$D_{JS} = \sqrt{\frac{1}{2}\left[D_{KL}(P_n(l), \frac{P_n(l)+P_c(l)}{2}) + D_{KL}(P_c(l), \frac{P_n(l)+P_c(l)}{2}) \right]}$$

$$D_{KL}(P,Q) = \sum_l P(l) \log_2 \left[\frac{P(l)}{Q(l)} \right]$$

$$0 \le D_{JS} \le 1$$

Jensen-Shannon Distance



JSD = 1.0
|dMML| = 0.7
|dNME| = 0.0

JSD mainly driven by difference in **MML**

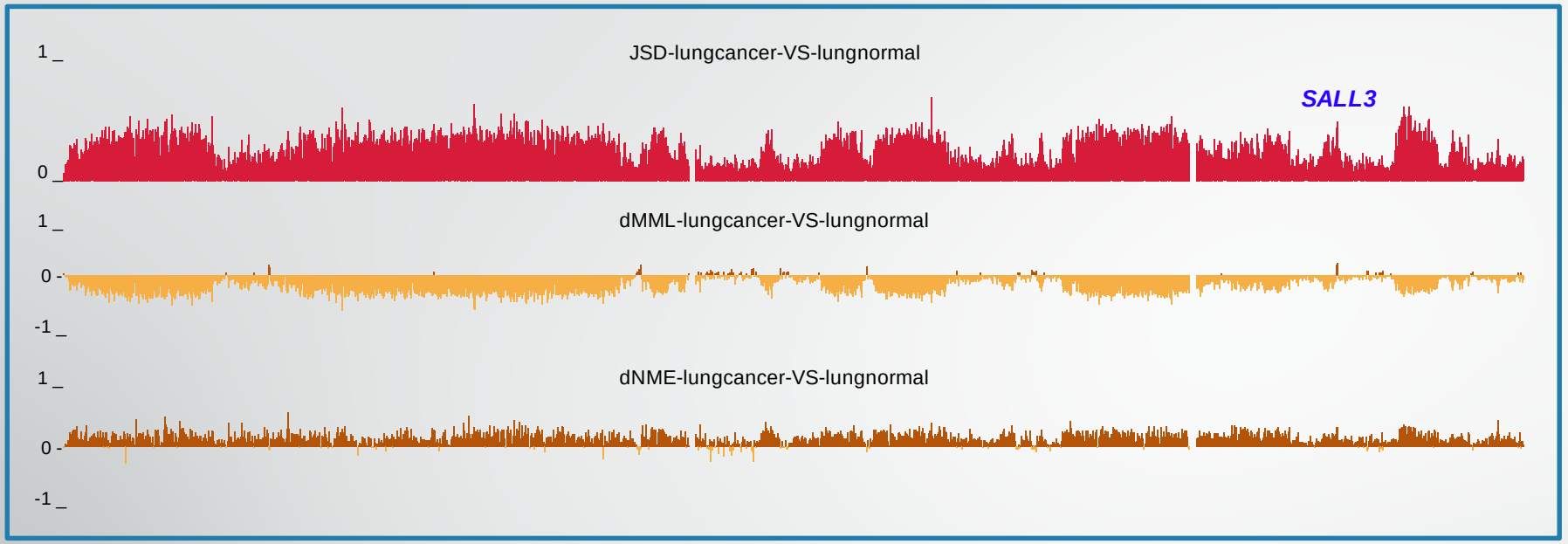
JSD = 0.9
|dMML| = 0.0
|dNME| = 0.6

JSD mainly driven by difference in **NME**

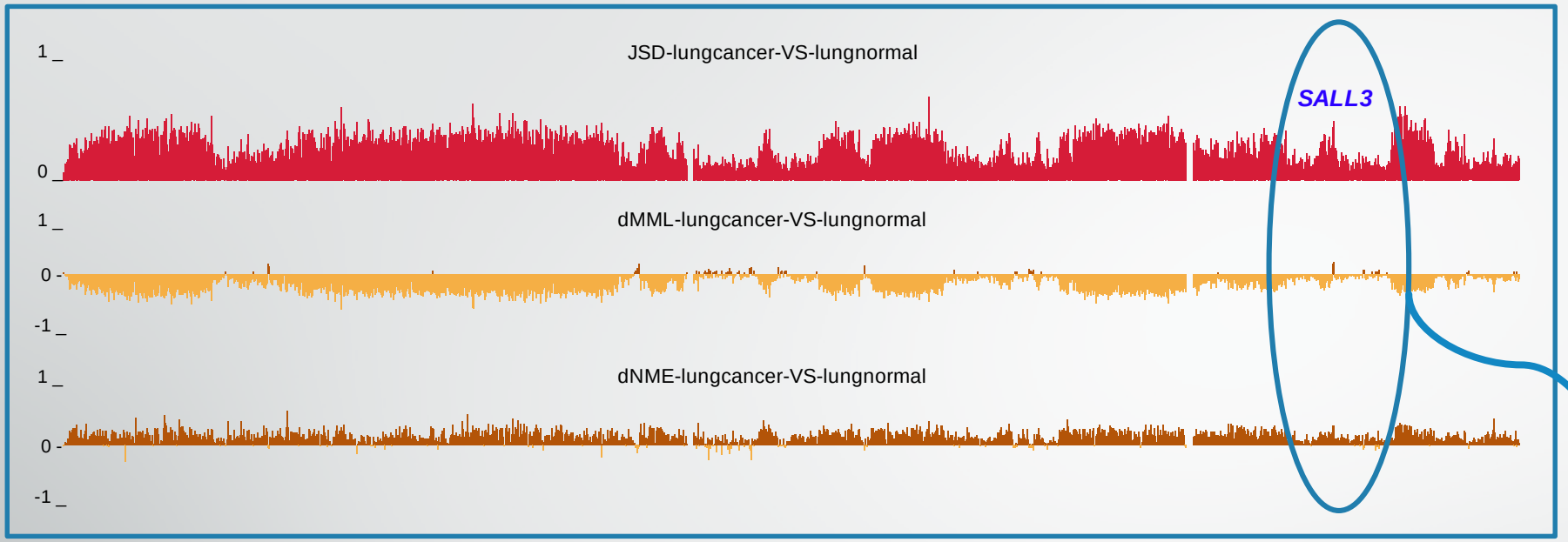
JSD = 1.0
|dMML| = 0.0
|dNME| = 0.0

JSD driven by factors **other than** MML or NME

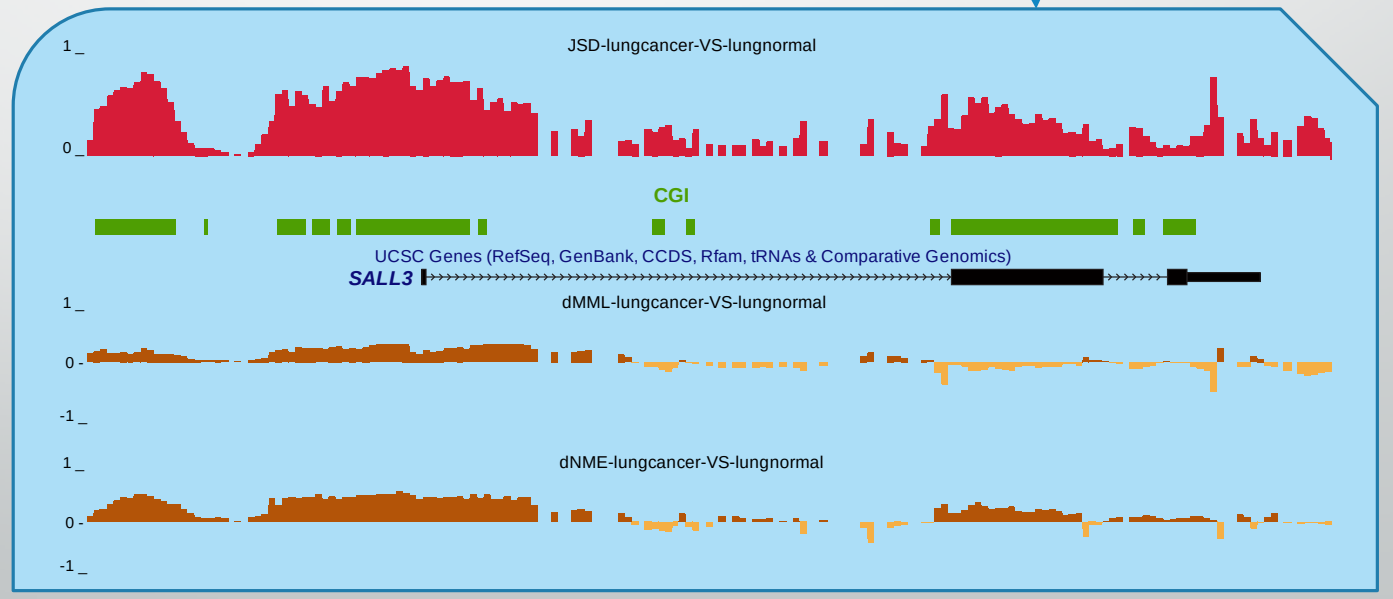
Jensen-Shannon Distance



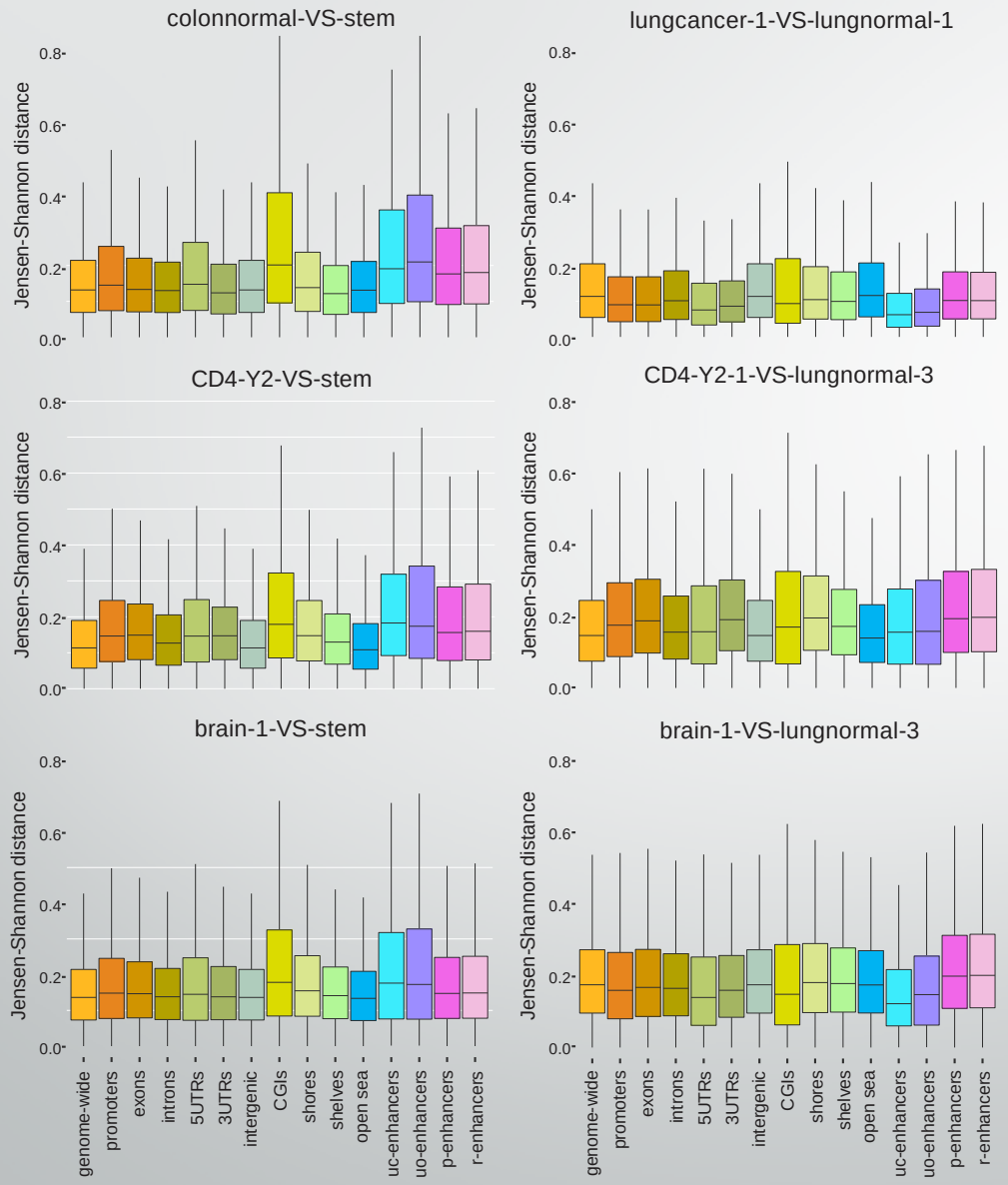
Jensen-Shannon Distance



SAL3 has been implicated in lung cancer

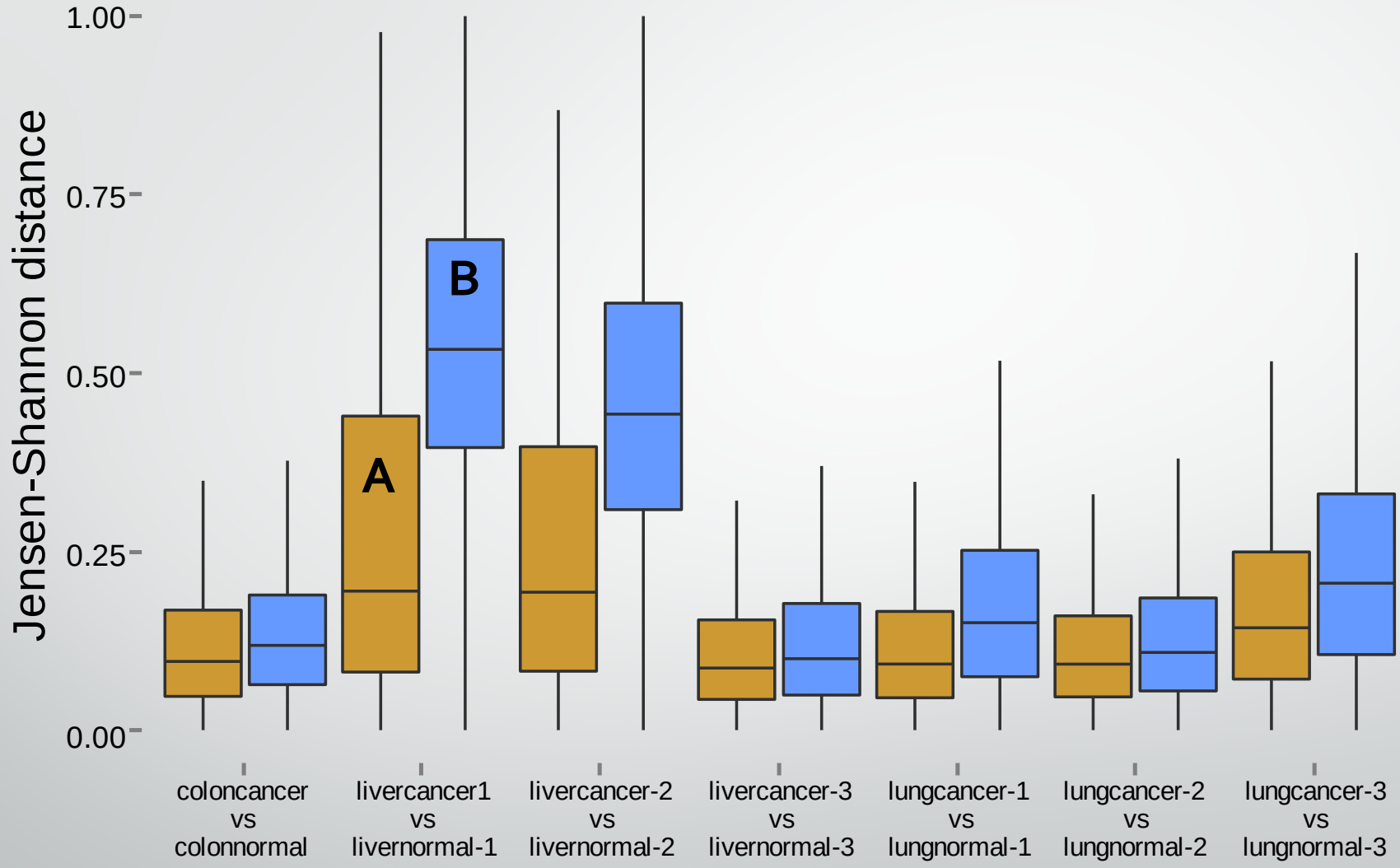


Jensen-Shannon Distance



No consistent containment of epigenetic dissimilarity within a particular genomic feature.

Jensen-Shannon Distance



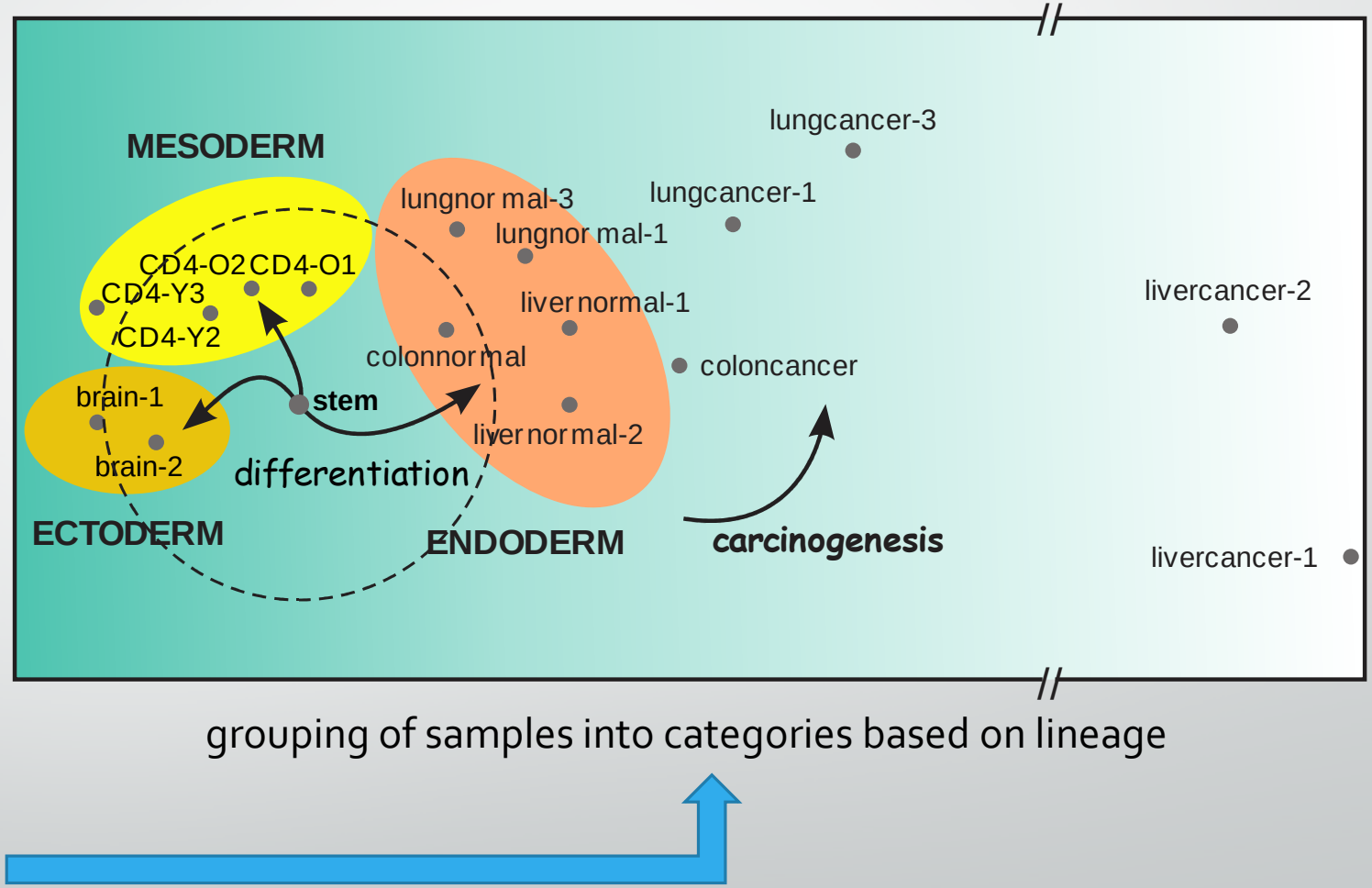
Larger epigenetic discordance is observed in compartment B than in A.

Jensen-Shannon Distance

- Quantified epigenetic discordance between two samples by calculating the average of all JSD values computed genome-wide.
- Computed this measure of epigenetic discordance between 17 representative samples.
- Formed the corresponding dissimilarity matrix.
- Used multidimensional scaling to find a 2D configuration of points whose inter-point distances approximately correspond to the epigenetic dissimilarities among the samples.

Jensen-Shannon Distance

- Quantified epigenetic discordance between two samples by calculating the average of all JSD values computed genome-wide.
- Computed this measure of epigenetic discordance between 17 representative samples.
- Formed the corresponding dissimilarity matrix.
- Used multidimensional scaling to find a 2D configuration of points whose inter-point distances approximately correspond to the epigenetic dissimilarities among the samples.



Sensitivity Analysis

- Epigenetic changes integrate environmental signals with genetic variation to modulate phenotype.
- Quantify influence of environmental exposure on methylation stochasticity.
- View environmental variability as a process that directly influences the parameters of the methylation potential energy landscape (Ising parameters).

Entropic Sensitivity Index

- Consider the case for which the Ising parameters $\boldsymbol{\theta}$ fluctuate around their true values by a random amount $G \times \boldsymbol{\theta}$, where G is a zero-mean normal distribution with standard deviation σ.
- Can show that

$$\sigma_h \simeq \eta \times \sigma$$

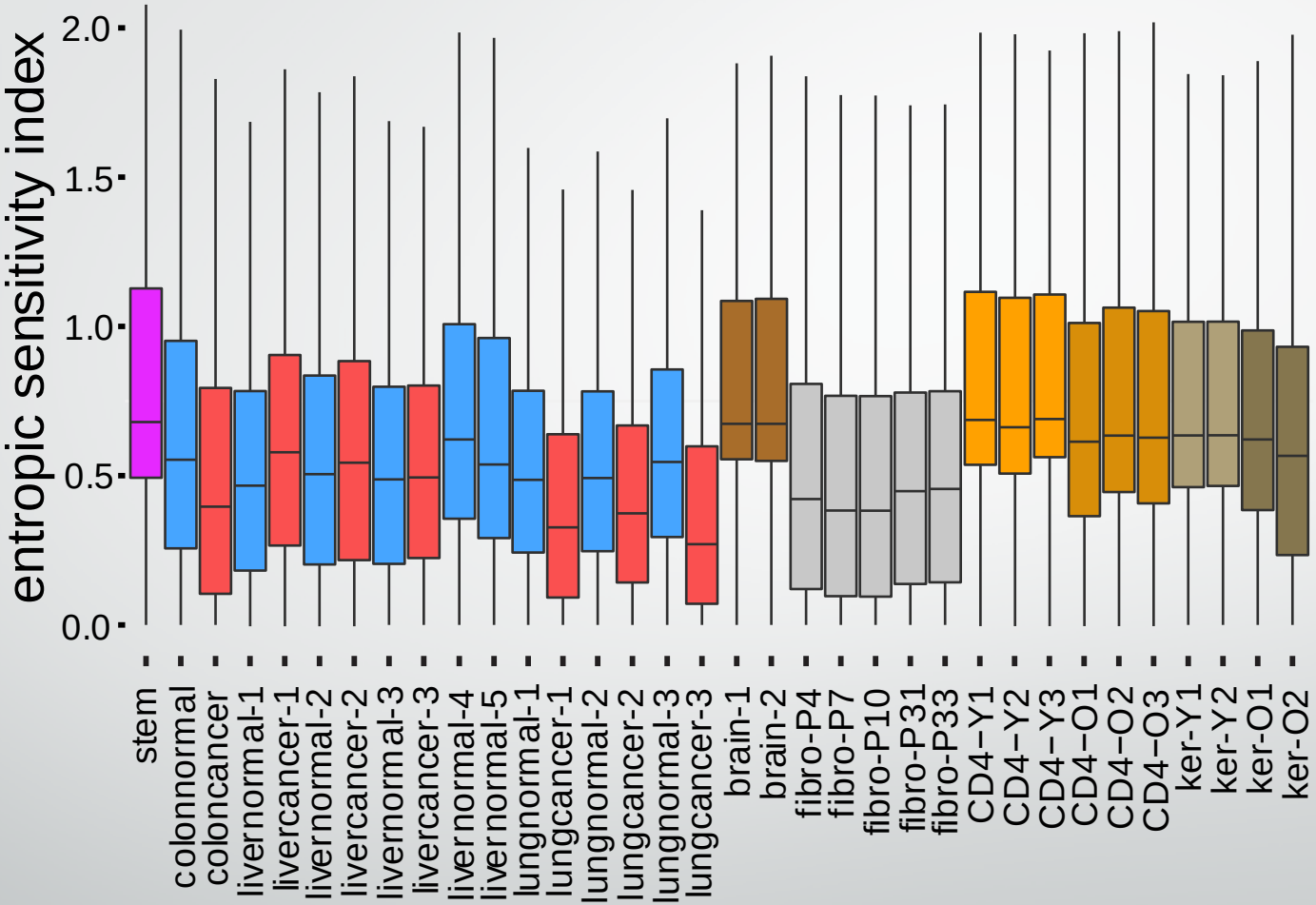
variation in entropy (NME)

environmental variation

$$\eta = \left| \frac{\partial h(g)}{\partial g} \right|_{g=0}$$

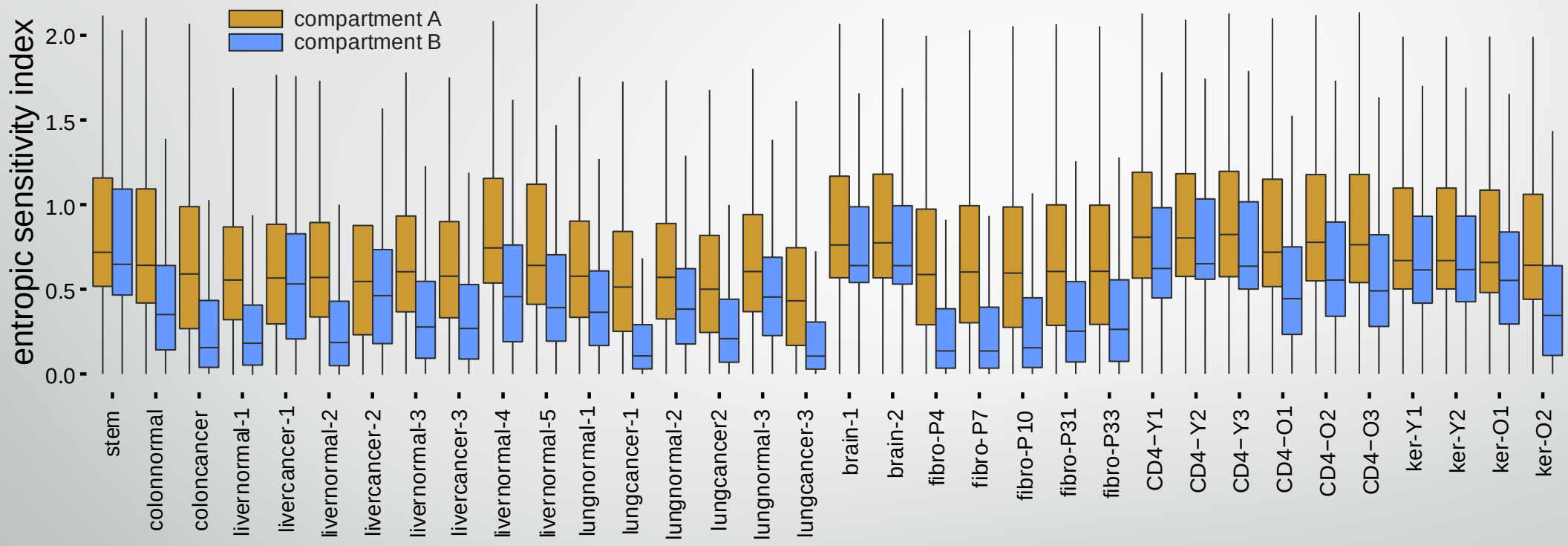
Entropic sensitivity index (ESI)

Sensitivity Analysis



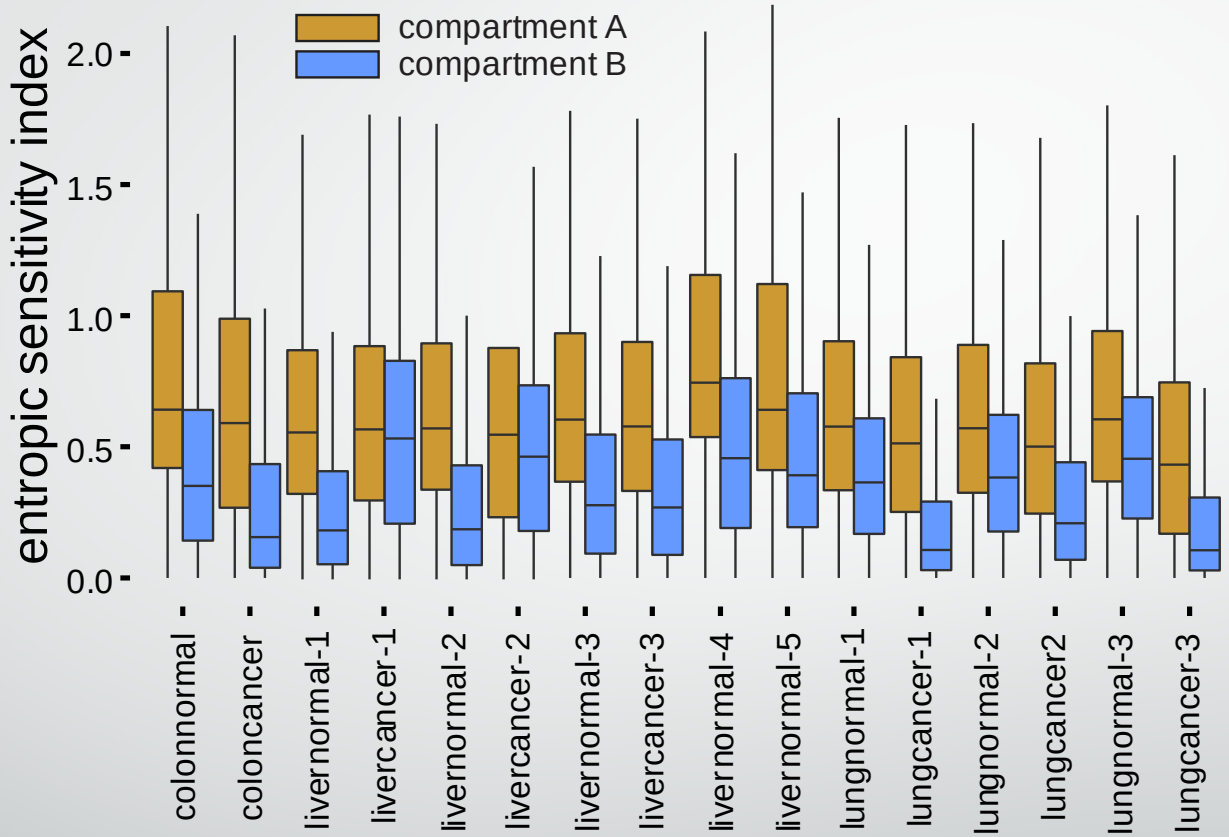
Globally observed differences in entropic sensitivity among tissues.

Sensitivity Analysis



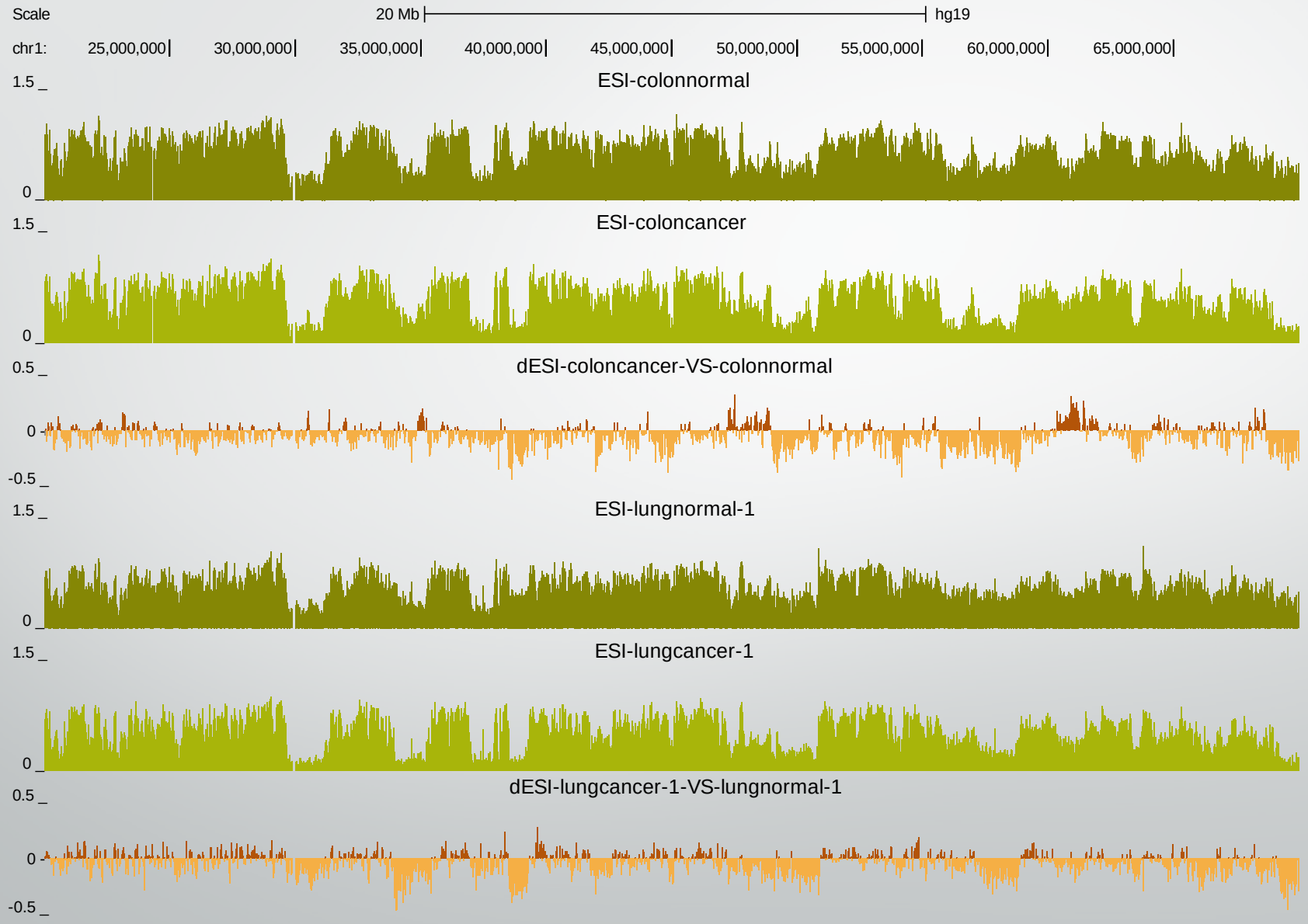
Entropic sensitivity within compartment A is higher than in compartment B.

Sensitivity Analysis



Observed differences between normal and cancer are largely confined to compartment B.

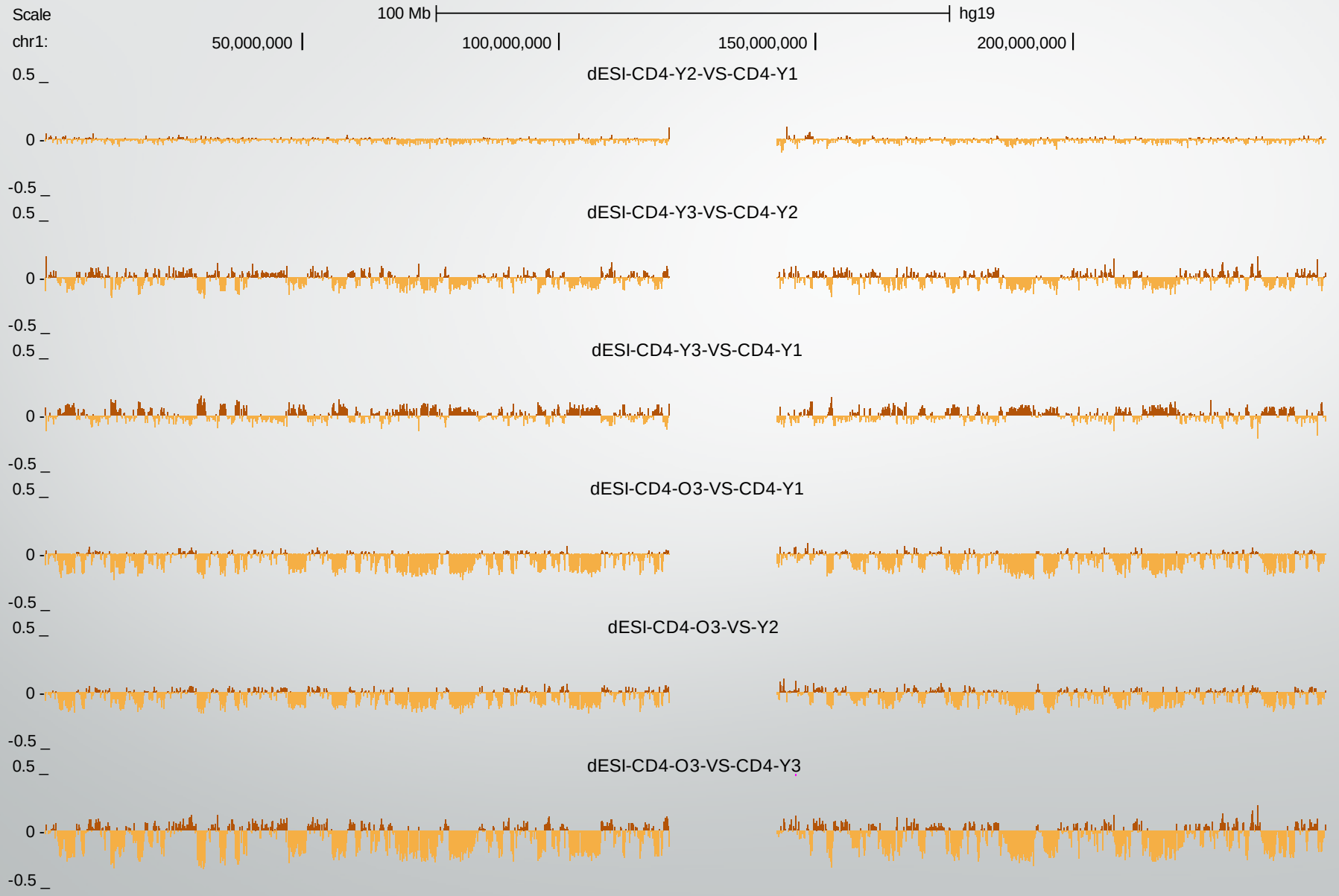
Sensitivity Analysis (Normal-Cancer)



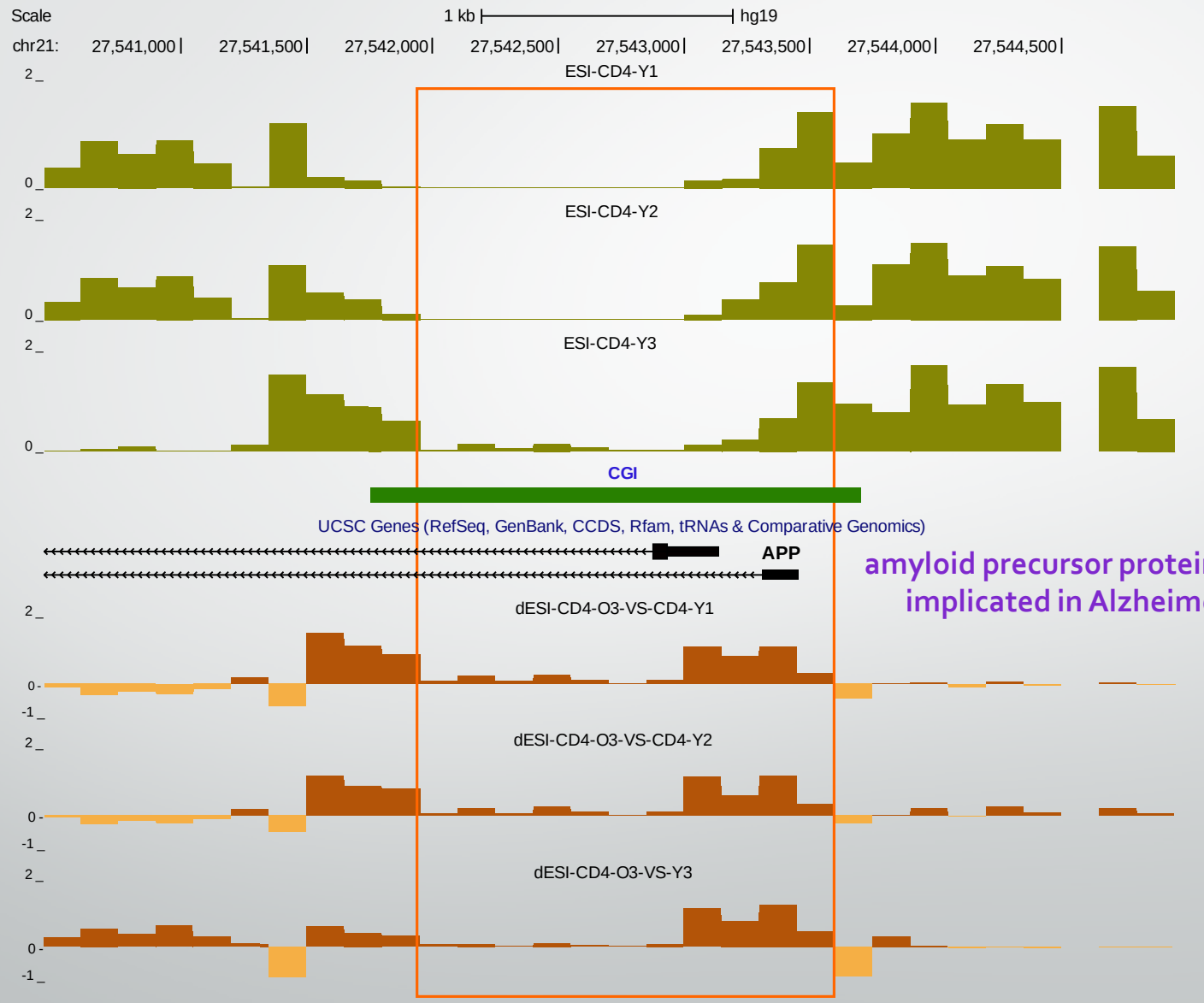
Sensitivity Analysis (Normal-Cancer)

polycomb target gene proposed as a candidate gene for type 2 diabetes

Sensitivity Analysis (Young-Old)



Sensitivity Analysis (Young-Old)



amyloid precursor protein gene implicated in Alzheimer's

Conclusion

- An information-theoretic approach to epigenomics based on the Ising model provides a formal foundation to methylation analysis.
- Yields fundamental insights into epigenetic behavior.
- Demonstrates a relationship between
  - chromatin structure
  - methylation channels
  - entropic sensitivity
- This may maximize an organism's efficiency in storing epigenetic information and help explain developmental plasticity.

Conclusion

- Pluripotent stem cells require relatively **high energy** to maintain **high capacity** methylation channels within a portion of the genome, achieving **reduced methylation stochasticity**.
- Regions characterized by **increased entropic sensitivity** are associated with **highly deformable potential energy landscapes**, which may correspond to differentiation branch points.
- After differentiation, some large genomic domains, such as regions associated with pluripotency, need not maintain high channel capacities and energy consumption
  - their sequestration providing **increased energy efficiency** with the cost of **high epigenetic stochasticity** and **reduced responsiveness**.